the basics of probability

- A sample space is the set of all possible outcomes of an experiment.
- An event is a subset of the sample space. Notice its a subset, not a single element (single outcome).
- we'll denote probability by P. probability of event A is $P(A)$
- a naive definition: $P(A) = \frac{num\ of\ favorable\ outcomes}{num\ of\ possible\ outcomes}$. The big assumption with this naive definition is that all outcomes are equally likely and that there is a finite number of possible outcomes. Obviously there are many cases where this assumption does not hold and in those cases, we can't use this definition.
- to calculate the numerator and denominator, we need to know how to count.
  - multiplication rule of counting: if we have r experiments with number of possible outcomes $n_1, n_2..n_r$, the combined experiment has $n_1 n_2 ... n_r$ possible outcomes.
  - the number of ways of picking groups of size k out of a set of n things is $\binom{n}{k}$ and is turns out to be $\frac{n!}{(n-k)!k!}$. Note that order doesn't matter. If order does matter, it is $P_k^n = \frac{n!}{(n-k)!}$. These are both quite easy to derive from the multiplication rule.
  - $\binom{n}{k}$ is called the binomial coefficient.
- there a four possible ways to sample: with and without replacement, and order matters and order doesn't matter.
  - order matters, with replacement: n^k
  - order matters, without replacement: P_k^n
  - order doesn't matter, with replacement:
    - if we have n things and we want to choose k of them, with replacement, it is as good as the number of non-negative integer solutions (0 allowed) of the equation $x_1 + x_2 + ...x_n = k$. To find this, we arrange this as $n + k - 1$ slots, and we choose k slots with without replacment. This turns out to be $\binom{n+k-1}{k}$.
  - order doesn't matter without replacement: $\binom{n}{k}$
- Labelling the items of a group of things is useful when doing counting problems and probability. This can be a number 1, 2, ...n. This is just an ID that makes things clearer to think about.
- Note that $\binom{n}{k} = \binom{n}{n-k}$
- A story proof in counting is a proof that uses a case or a story to prove something instead of algebra, etc. examples:
  - to prove $\binom{n}{k} = \binom{n}{n-k}$, we can say something like the number of ways of picking k things from n things is the same as the number of ways of picking (n-k) things from n things because when you pick n things you implicitly also pick n-k things.
  - to prove that $n\binom{n-1}{k-1} = k\binom{n}{k}$, think of the number of ways of picking k people out of n with one designated as the tech lead. there two approaches here, and the LHS and RHS correspond to those approaches.
  - $\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$ (this is called Vandermonde's identity). the story is pretty obvious.

  these stories are just counting the same thing in different ways. if both ways are correct, they must be equal. notice that in both cases above, its also pretty easy to give a proof from the definition just using algebraic manipulation.
- Disjoint sets are basically events with no intersection. note that an event is a subset of the sample space (set of all possible outcomes). notice in the last identity above, across iterations, we have disjoint sets. We can't just add them up if they are not disjoint sets (if they have overlap. in other words, if intersection size isn't 0).

- general definition of probability (without the assumption of equal likelihood) of outcomes:
  A probability space consists of S and P. S is the sample space (which is already defined except we assumed that it is finite but now we'll break that assumption). So set S need not have a finite size.
  P is a function whose domain is all subsets of set S. In our terminology, P takes in an event. P gives out a number in [0,1] such that:
  - $P(\phi) = 0, P(S) = 1$ where ϕ is the empty set and S is the as defined above. Impossible events have probability 0 and any possible event has probability 1.
  - $P(U_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ whenever $A_1, A_2, ...A_n$ are disjoint sets.

  That's it. This is the definition, also called axioms of probability. Herb Gross might call it the "structure".
- Every single theorem in probability follows from these two simple rules. Notice how the intuition of fairness is not at all encoded into this definition. That's surprising to me. I'm curious to find out why. From this definition, I can say that the probability of a "fair" coin is defined such that P(head) = 0.2 and P(tail) = 0.8 or I can say that P(head) = 0.3 and P(tail) = 0.7 and these are both valid by the above definition.
- birthday problem. if we have k people in a room, what's the chance we have at least 1 pair of people who have the same birthday? first of all, if $k > 365$, then probability is 1. think of it as the balls and buckets problem. this is called the *pigeonhole principle*. if we have more balls than boxes, at least 1 box will have more than 1 ball.
  if $k \leq 365, P(no\ match) = \frac{365*364...(365-k+1)}{365^k}$.
  P(match) is 1 - P(no match). Turns out if for this number to be >50% we just need $k \geq 23$ people in the room.
- the above is similar to probability of collision in hashing buckets etc.
- useful strategy: to find the probability of something see if finding the probability of the complement of that event helps.
- "the biggest coincidence is when there are not coincidences. there are a mind boggling number of coincidences that could occur in the world so it must be unlikely that none of them occur".
- Properties that follow from the definition (axioms) of probability:
  - $P(A^c) = 1 - P(A)$ → this follows from the fact that A and A^c are disjoint
  - if $A \subseteq B$, then $P(A) \leq P(B)$.
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. To prove this, we split $A \cup B$ into 2 disjoint sets and we can apply the axioms. (double counting)
  - Similarly we can show
    $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$
    The generalization of this to many more events than just 2 or 3 is called the inclusion-exclusion principle. Its just a matter of subtracting and adding until you make sure all regions are counted just once and there's no over-counting or under-counting:

    $$p(a_1 \cup a_2 ... \cup a_n) = \sum_{j=1}^{n} p(a_j) - \sum_{i<j} p(a_i \cap a_j) + \sum_{i<j<k} p(a_i \cap a_j \cap a_k) - (-1)^{n+1} p(a_1 \cap ... \cap n_n)$$

  - to prove this, we basically need to make sure that each area in a venn diagram with n intersecting circles only gets counted once. we can show that if an area is in exactly k of n circles, its count is affected by $\binom{1}{k} - \binom{2}{k} + \binom{3}{k}$.... Which we can prove is 1 by using the idea from binomial expansion for $(x-y)^n$ when $x = 1, y = 1$.
  - the inclusion exclusion principle is a pretty powerful general technique for finding the probability.