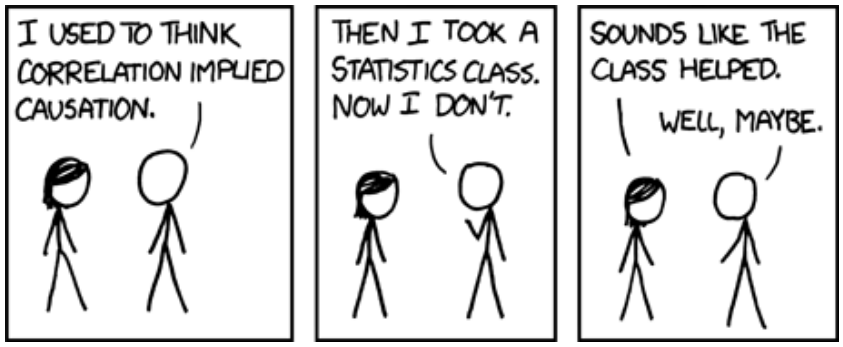
CS 109 NOTES

ARUN DEBRAY
JUNE 11, 2013

These notes were taken in Stanford's CS 109 class in Spring 2013, taught by Professor Mehran Sahami. I live-TEXed them using `vim`, and as such there may be typos; please send questions, comments, complaints, and corrections to `adebray@stanford.edu`. Thanks to Rebecca Wang and Shivaal Roy for catching a few errors.

Contents



1. Counting: 4/1/13

"One, ah, ah, ah! Two, ah ah ah! Three, ah ah ah!" – Count von Count, Sesame Street

Though that average CS student might wonder why he is learning probability, it is in fact an incredibly useful way of understanding principles and techniques in CS. The traditional view of probability (placing balls in urns) isn't as helpful, but there are a lot of problems in CS which are just abstractions of the conventional view of probability.

For example, how does Amazon determine recommendations of similar orders? How does a spam filter classify messages? Both of these applications rely on probability.

Here's a list of other applications:

(1) In terms of algorithms, there is a whole class called randomized algorithms. Additionally, analysis of deterministic algorithms can be aided by probabilistic analysis and learning the expected value.
(2) Hashing and sorting (especially Quicksort) both deeply rely on randomness.
(3) Systems design also depends on randomness.
(4) Machine learning depends heavily on statistical models. And of course machine learning has huge applications, such as financial modeling, DNA analysis, etc.

But to get to all of these applications, it will be necessary to start from the bottom. The *very* bottom. Let's learn how to count.

An experiment is some action that has a set of possible outcomes. Conducting the experiment indicates one of the outcomes, but it's not possible to predict the outcome beforehand (e.g. flipping a coin). An event is some subset of outcomes we might care about (e.g. rolling a D6 and getting an even number). How can we count the number of outcomes in an event?

The first rule to learn is the sum rule of counting: if the outcomes of some experiment can come from two sets A and B, such that $|A| = m$, $|B| = n$, and $A \cap B = \emptyset$, then the number of outcomes is $m + n$.

Example 1.1. This is as easy as it sounds: suppose Facebook has 100 machines in San Jose and 250 machines in Denver. Then, if someone sends a login request, there are 350 possibilities for which machine that request could go to.

The next rule is the product rule of counting: if an experiment has two parts that don't depend on each other (e.g. rolling two dice), and the outcome of the first part is from a set A, where $|A| = m$ again, and the outcome of the second part comes from a set B, with $|B| = n$, then the total number of outcomes is mn.[1]

There are a couple of functions you should know because they come up a lot in probability:

Definition.

- The floor function of x is $\lfloor x \rfloor$, which is the largest integer less than or equal to x. This is just rounding down: $\lfloor 1/2 \rfloor = 0$, $\lfloor 2.9 \rfloor = 2$, and $\lfloor 8 \rfloor = 8$. Be careful with negative numbers: $\lfloor -1/2 \rfloor = -1$.
- The ceiling function of x is $\lceil x \rceil$, which is the smallest integer greater than or equal to x. Thus, $\lceil 1/2 \rceil = 1$, $\lceil 8 \rceil = 8$, and $\lceil 2.9 \rceil = 3$.

  These functions are used, for example, when it is necessary to fit an integer value into something given by a real number, such as when an optimization calls for 2.9 people or something similarly problematic.

The inclusion-exclusion principle is slightly more general than the sum rule: if the outcome of an experiment is given from sets A and B (and we don't know anything about the intersection of A and B), then the number of outcomes is $|A \cup B| = |A| + |B| - |A \cap B|$.

Example 1.2. As a CS major, you are familiar with binary numbers.[2] How many 4-bit binary sequences start or end with a 1? It turns out there are 12 such possibilities: there are 8 strings that start with a 1, 8 that end with a 1, and 4 strings that start and end with a 1. Thus, the total number is $8 + 8 - 4 = 12$ such strings.

Another useful thing to know is the General Pigeonhole Principle: if $m, n \in \mathbb{N}$ and m objects are placed in n buckets, then at least one bucket contains at least $\lceil m/n \rceil$ objects.[3] For example, if a hash table has 100 buckets and 950 strings hashed into it, then it can be guaranteed that at least one bucket has 10 or more strings in it. We cannot guarantee that there would be a bucket with 11 strings in it, and an explicit counterexample can be given: 50 buckets have 10 strings each in them, and 50 buckets have 100 strings each in them. Of course, the individual buckets could have anything in them, or could be empty.

---

[1] Hey, it's Theorem 13.27 from Math 121!

[2] But it is possible to drink binary numbers, too, thanks to the existence of Pepsi One and Coke Zero. This would make for a very slow, but very refreshing, Turing machine.

[3] The name denotes pigeons nesting in their pigeonholes, but there's an alternate formulation: if one drills m holes in n pigeons, with $m > n$, then at least one pigeon has at least one hole in it!

Time for yet another principle, called the General Principle of Counting. This says that if r experiments are performed such that part 1 of the experiment has n_1 outcomes, part 2 has n_2 outcomes, and so on until part r has n_r outcomes, then the total number of outcomes is $n_1 n_2 \dots n_r$.

Example 1.3. California license plates were once given by 3 uppercase letters, followed by 3 digits. Thus, the total number of distinct license places was $26^3 \cdot 10^3 = 17,576,000$.

Of course, this is less than the number of cars in California, so this was recently changed to add another digit, so there are now enough letters. And if you think this is bad, consider IP addresses.

Definition. A binary function f is a function $f : \{1, \dots, n\} \to \{0, 1\}$.

Thus, there are 2^n possible binary functions on $\{1, \dots, n\}$. This can be shown by realizing that each binary function corresponds to a unique binary string of length n, and vice versa, and there are 2^n such strings.

This illustrates a common theme of this class: the math isn't all that difficult, but what's tricky is understanding how to turn the problem into a mathematical question.

Definition. A permutation is an ordered arrangement of distinct objects.

There are $n!$ possible permutations of n distinct objects. This is because the first slot can be filled by all n objects, then the second has $n - 1$ possibilities (since one is already in a spot), then $n - 2$, etc.

This can be used to make statements in which orders are restricted. If one has 4 PCs, 3 Macs, and 2 Linux boxes, one could try to order them in a way such that all computers of a given OS are adjacent. The total number of possibilities is $4!3!3!2!$: within each group, all machines of a given OS can be permuted, giving $4!$, $3!$, and $2!$, and these are unrelated events, so they are multiplied together. Then, these groups can be permuted, giving another $3!$

2. Permutations and Combinations: 4/3/13

> "Should array indices start at 0 or 1? My compromise of 0.5 was rejected without, I thought, proper consideration." – Stan Kelly-Bootle

We're encouraged to use LaTeX in this class, which is exciting!

Permutations, discussed in the previous lecture, can be applied to binary search trees.

Definition. A binary search tree (BST) is a binary tree such that for every node n:

- $n \geq \ell$, where ℓ is the value of its left child;
- $n \leq r$, where r is the value of its right child;
- The left and right subtrees are both binary search trees.

You might remember that the insertion time is about $O(\log n)$.

How many degenerate BSTs of length 3 (for 3 distinct elements) are there? These are the BSTs in which each node has at most one child. There are $3!$ ways to order the elements to be inserted. Note that depending on order, the placement of left or right children will change: in $1 - 2 - 3$, there will be two right children, but in order $2 - 1 - 3$, one obtains both left and right children of the element 2. Thus, $2 - 1 - 3$ and $2 - 3 - 1$ aren't degenerate, but the rest are, so there are four such degenerate trees. When the size of the tree increases, the probability that a randomly selected tree is degenerate becomes vanishingly small.

Here's another problem: given 2 0s and 3 1s, how many different strings of length 5 can be made? This is more interesting than the previous problems, as some of the permutations are identical. Thus, distinguish the zeros and ones, so that each option is distinct, calling them $0_1, 0_2$, etc. Thus, there are $5! = 120$ orderings of the distinct elements. However, any permutation of the zeroes is unimportant in the final answer, and similarly for the ones, so the final result is $5!/2!3! = 10$.

For another example, consider the RCC who works on 4 PCs, 3 Macs, and 2 Linux boxes, and we consider two machines to be identical iff they have the same OS. Then, the number of distinct orderings is $9!/2!3!4! = 1260$.

Okay, now for some security: consider iPhone passwords. These are 4-digit passwords, so there are 10000 possibilities for passwords. However, it's possible to obtain the four digits of the password based on looking at where grease smudges exist on the screen. This reduces the search space to $4! = 24$, which is much less secure. What if there are only 3 smudges? Then, it's not obvious which one is repeated. If the digits are a, b, and c, without loss of generality assume c is repeated (so we have to multiply the final answer by 3). Then, this is a problem we know how to solve: $4!/2! = 12$, and then factoring in the chance that a or b might be doubled gives $12 \cdot 3 = 36$ options. Slightly more secure, interestingly.

But then, two digits must be better, right? Then, suppose the two digits are each given twice: the computation is $4!/2!2! = 6$, but there are no other possibilties. If one digit is used three times, the computation is $4!/3! = 4$, but there are two possibilities, so there are in total $6 + 8 = 14$ options. This isn't as secure.

The general formula is: if one has n objects and there are distinct groups of objects that are considered identical with sizes $n_1, \dots, n_r$, then the number of distinct permutations is $n! / \prod_{i=1}^{r} n_i!$.

Next is (of course) combinations. This is an unordered selection of r objects from n objects. This can be modelled by ordering all n objects, and then treating the first r as identical and the last $n-r$ to be identical in the above formula. Thus, we obtain $\binom{n}{r} = n!/(r!(n-r)!)$, sometimes also written $C(n,r)$. It has the important property that $\binom{n}{r} = \binom{n}{n-r}$.

Example 2.1. Mehran has six big books on probability (including the modestly titled *All of Statistics*). If you wanted to borrow three of them, there are $\binom{6}{3} = 20$ options. However, supposing two of them are different editions of the same book, one might not want to get both of them. Under this constraint, break the problem into three subproblems:

(1) Suppose the 8^{th} edition and two other books are chosen. Then, there are $\binom{4}{2}$ options, since the 9^{th} edition can't be picked.
(2) Symetrically, getting the 9^{th} edition yields another $\binom{4}{2}$ choices.
(3) Neither edition could be selected, which gives $\binom{4}{3}$ choices.

Thus, adding them together gives the total number: 16. Dividing this into subproblems is a good strategy, but it's important to make sure they cover the whole space and don't overlap.

Alternatively, you could count all the options that you don't want: there are four such options, given by both editions and another book. Thus, there are $20 - 4 = 16$ options, again.

The formula for $\binom{n}{k}$ has a nice recursive definition which can be illustrated with some code: pick some special point. Either it is in the combination, or it isn't, and these are all of the distinct cases. Thus, there are $C(n-1,k-1)$ options if it is included, and $C(n-1,k)$ if it isn't included. This is nice, but recursion usually needs the base cases: $C(n,0) = C(n,n) = 1$.

Here's how this looks in code:

```
int C(int n, int k) {
    if (k == 0 || n == k) return 1;
    return C(n-1,k) + C(n-1,k-1);
}
```

$\binom{n}{k}$ is what is called a binomial coefficient: it is named so because it appears in the formula given below. There are two proofs in the textbook, but they're not all that important in the class. However, the equation is very important.

Theorem 2.1 (Binomial).

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

For an application, consider a set with n elements. Then, what is the size of the power set? If one chooses a subset of size k, there are $\binom{n}{k}$ such options, so the total number of subsets is

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} = (1+1)^n = 2^n.$$

More generally, one has the multinomial coefficient $\binom{n}{n_1,\dots,n_r} = n!/n_1! \cdots n_r!$, where $n_1 + \cdots + n_r = n$. Technically, this is overdetermined, and sometimes the n_r-term is dropped (since it can be recovered).

Example 2.2. Suppose Google has 13 machines and 3 data centers and wants to allocate machines to data centers A, B, and C which can hold 6, 4, and 3 machines, respectively. Then, there are $13!/6!4!3! = \binom{13}{6,4,3}$ possibilities.

Returning to the ball-and-urn formalism, how many ways are there to place n distinct balls into r urns? There are r^n possibilities, since any ball can go into any urn.

If they are taken to be indistinct, but the urns aren't, then the question becomes a bit more nuanced. There are n balls and r urns, but once everything is placed the indistinctness of the balls means there are $n+r-1$ distinct objects: in some sense, the dividers between the urns are what we care about. Thus, the number of options is $(n+r-1)!/n!(r-1)! = \binom{n+r-1}{r-1}$. The total number of options is divided by the number of ways the balls can be permuted and the dividers can be permuted. This is because the urns can be represented as dividers.

In CS, there is a slightly different example: consider a one-indexed array `x` of length r such that $0 \le \texttt{x[i]} \le n$ for each `i`, and such that the sum of the elements is n. This is a reformulation of the balls problem: each index of the array is an urn, and its value represents the number of the balls in the urn. The book calls this 6.2, and has a related proposition 6.1:

Proposition 2.2. *Consider the slightly related problem in which every urn has at least one ball in it.*[4] *This can be solved by giving each child one piece of candy, which gives a slightly different spin on the previous problem: there are $n-r$ candies to put in n baskets. Thus, the total number of options is $\binom{(n-r)+(r-1)}{r-1} = \binom{n-1}{r-1}$.*

[4]Real-world example: trick-or-treating. You wouldn't want some kid to go home candyless, would you?

Consider some venture capital firm which is required to distribute \$10 million amongst 4 distinct companies A, B, C, and D, in \$1-million increments. Thus, the goal is to put 10 balls in 4 urns, and some company might be left out in the cold, giving $\binom{10+4-1}{4-1} = \binom{13}{3} = 286$.

But if you want to keep some of it, the answer is $\binom{14}{4}$, since you can be thought of as adding on more company: your own back account. This is much nicer than all the casework that is the obvious approach.

Lastly, suppose that A only will accept payments of at least \$3 million. This reduces the search space: there are $\binom{10}{3}$ choices, because after allocating \$3 million to A, there is \$7 million left to put in 4 companies, giving $\binom{10}{3}$. Add this to the case where A is given no money, which is an exercise in allocating \$10 million to 3 companies.

This final question used to be used at company A (i.e. Google). Consider an $n \times n$ grid with a robot in the lower left-hand corner of the grid and a destination in the right-hand corner. If the robot can only move up or right, it will reach the destination, but the path isn't unique. How many distinct paths are there? The robot needs to make $n-1$ moves up and $n-1$ moves right, so the goal is to find the number of permutations of these moves given that the moves in the same direction are identical: $(2n-2)!/(n-1)!(n-1)! = \binom{2n-2}{n-1}$.

3. Probability: 4/5/13

"I like to think of the size of a set as how much Stanford spirit it has: its cardinality."

Today's lecture will help with the material in the second half of the first problem set. However, in order to understand probability some of the theory behind it will be necessary.

Definition. A sample space is the set of all possible outcomes of an experiment.

Example 3.1. Some examples of sample spaces:

- When flipping a coin, the sample space is $S = \{H, T\}$.
- When flipping two distinct coins, the sample space is $S \times S = \{(H,H),(H,T),(T,H),(T,T)\}$.
- When rolling a six-sided die, the sample space is $\{1,2,3,4,5,6\}$.
- Sample spaces need not be finite: the number of emails sent in a day is $\{0,1,2,3,\dots\}$.
- They can also be dense sets: the number of hours spent watching videos on Youtube in a day instead of working is $S = [0,24]$.

Definition. An event, usually denoted E, is a subset of the sample space.[5] This is in some sense the outcome we care about.

Example 3.2. Here are some events corresponding to the sample spaces in Example 3.1:

- If a coin flip is heads, $E = \{H\}$.
- If there is at least one head in two flips, $E = \{(H,H),(H,T),(T,H)\}$.

Now let's take some set operations. If $E, F \subseteq S$ are events in the sample space S, then an event that is in E or F is represented by $E \cup F$, and an event that is in E and F is given by $E \cap F$.[6] For example, if E represents rolling a 1 or a 2 on a die and F represents rolling an even number, then $E \cap F$ is rolling a 2.

Definition. Two events E, F are called mutually exclusive if $E \cap F = \emptyset$.

One can take the complement of an event E, $\sim E = S \setminus E$, or everything that isn't in E. The book uses E^{C}. This leads to De Morgan's Laws: $\sim(E \cup F) = (\sim E) \cap (\sim F)$ and $\sim(E \cap F) = (\sim E) \cup (\sim F)$. This can be inductively generalized to n events in the unsurprising manner.

Probability is given in several different ways. The first will be given now, and another will be given later. The frequentist interpretation of probability interprets the probability as the relative frequency of the event. Let $n(E)$ be the number of times event E happens when the experiment is done n times; then, the probability is $P(E) = \lim_{n\to\infty} n(E)/n$. With this definition comes some axioms:

(1) Probabilities must lie between 0 and 1.
(2) $P(S) = 1$.
(3) If E and F are mutually exclusive events, then $P(E) + P(F) = P(E \cup F)$.

From axiom 3 most of the interesting mathematics comes up. For just one example, the multivariate case is a straightforward generalization: if $E_1, E_2, \dots$ is a sequence of mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

[5] The textbook writes that $E \subset S$, but $E = S$ is allowed too, so be careful.

[6] More questionable textbook notation: the textbook uses EF to denote $E \cap F$.

This can be generalized to the uncountable case, but things get weird, so take care.

There are several other elementary consequences of the axioms. They are easy to prove by doing a bit of set chasing or drawing Venn diagrams.

- $P(\sim E) = 1 - P(E)$.
- If $E \subseteq F$, then $P(E) \leq P(F)$.
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

The latter formula can generalize to n variables as

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P\left(\bigcap_{j=1}^{r} E_{i_j}\right).$$

This seems pretty scary, but all that it does is take the intersections of subsets of size r for all r in 1 to n, and adds and subtracts things based on the number intersected (so nothing is double-counted). All of the odd intersections have a positive coefficient, and the even ones have a negative coefficient. The $i_1 < \cdots < i_r$ means that every combination is considered, and order is used to ensure that such each combination is taken only once. For example, if $n = 3$,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C).$$

By drawing a Venn diagram, it should be possible to see why this works.

Some sample spaces have a property called equally likely outcomes. This means that each point in the sample space has the same probability: if $x \in S$, then $P(x) = 1/|S|$. Examples of these include fair coins and fair dice. Thus, if $E \subseteq S$, then $P(E) = |E|/|S|$.

Example 3.3. If one rolls two six-sided dice, the sample space is $S = \{1, \ldots, 6\}^2$, and if one wants to calculate the probability that the sum is 7, the event is $E = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$, so the probability is $|E|/|S| = 6/36 = 1/6$.

Example 3.4. Suppose one has four Twinkies and three Ding Dongs.[7] What is the probability that one Twinkie and 2 Ding Dongs are drawn if three are picked? There are two ways to look at this:

- One can consider it to be ordered: the Twinkie is either the first, second, or third item, giving $(4)(3)(2)$, $(3)(4)(2)$, and $(3)(2)(4)$ combinations respectively. Thus, the probability is $((4)(3)(2)+(3)(4)(2)+(3)(2)(4))/210 = 12/35$.
- Alternatively, it can be unordered: $|S| = \binom{7}{3} = 35$, and $|E| = \binom{4}{1}\binom{3}{2} = 12$ (picking the Twinkies and Ding Dongs respectively).

The takeaway is that either method works, but be careful not to accidentally mix them.

Example 3.5. Suppose someone is manufacturing chips, and has n chips, one of which is defective. If n is large, it's impractical to test them all, so only k will be examined. Thus, the sample space has $\binom{n}{k}$ possible choices, and the event has a size of $\binom{1}{1}\binom{n-1}{k-1}$ (since the bad chip is picked, and the rest are fine). Thus, after cancelling some terms, the probability is n/k.

Alternatively, here's a more intuitive formulation: pick k chips out of the n, and then randomly make a chip defective. Then, the chance that the chip was part of the original k is k/n. Intuition and cleverness isn't always necessary, but it's a great way to save time.

Example 3.6. Let's play poker! Suppose one has a 5-card poker hand drawn from a standard deck and wants a straight. This is a set of 5 cards of consecutive rank, e.g. $\{2, 3, 4, 5, 6\}$. Again, this is slightly different from the textbook's definition. The probability is a little weird, because there are lots of cases: $\{A, 2, 3, 4, 5\}, \{2, 3, 4, 5, 6\}, \ldots, \{10, J, Q, K, A\}$. The sample space has size $\binom{52}{5}$, and each of these has $\binom{4}{1}^5$ options. Thus, $|E| = 10\binom{4}{1}^5$, giving a probability of just under one percent.

But an "official" straight in poker is slightly different: a straight flush (i.e. all five cards are the same suit) is not considered to be a straight. There are $10\binom{4}{1}$ possible straight flushes, so $|E| = 10\binom{4}{1}^5 - 10\binom{4}{1}$ now.

Example 3.7. Consider flipping cards until an ace comes up, and take the next card. Is it more likely that the ace of spades is this next card, or the two of clubs? Here, $|S| = 52!$, since the deck is shuffled. The first case, in which the ace of spades is removed, and then reinserted after the first ace, there are $51! \cdot 1$ ways for this to happen (since the other cards can be anywhere). Thus, the probability is $51!/52!$.

The two of clubs has the same thing going on: there's one place to put it, so the probability is *the same*. This is counterintuitive.

[7] Dr. Sahami actually had these with him. How long has it been since Hostess stopped producing them again?

Example 3.8. Suppose 28% of students at a university program in Java, 7% program in C++, and 5% program in both. Then, what percentage of students program in neither?

When answering a problem, define an event very precisely, as this will make errors less likely. So let A be the probability that a randomly chosen student programs in Java and B be the probability that a student programs in C++. Thus, the probability that someone programs in neither is $1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A \cap B)) = 1 - (0.28 + 0.07 - 0.05) = 0.7$, so the answer is 70%.

What percentage of programmers use C++, but not Java? This is $P((\sim A) \cap B) = P(B) - P(A \cap B) = 0.07 - 0.05 = 0.02$, or 2%.

Notice how similar this looks to a problem on this week's homework.

One important consequence is the birthday problem. What is the probability that in a room of n people, no two share the same birthday? Then, $|S| = 365^n$ and $|E| = (365)(364)\cdots(365 - n + 1)$. Clearly bad things happen if $n \geq 365$. More interestingly, at 23 people, the probability that two people share a birthday is at least $1/2$. This is much less than intuition would expect. At 75 people, it's over 99%.

Now consider the probability that of n other people, none of them share a birthday with you? Here, $|S| = 365^n$ again, and $|E| = 364^n$. Thus, the probability is $(364/365)^n$, and at $n = 23$, the probability that nobody's birthday matches yours is about 94%. Even at 160 people it's only about 64%. In order to get up to a 50% chance, you need about 253 people. That's again quite different from intuition! The interaction of many-to-many and one-to-many effects is tricky.

4. Conditional Probability: 4/8/13

First, it will be helpful to understand the notion of distinctness and indistinctness and ordered and unordered sets. In many problems, it helps to temporarily view indistinct objects as distinct ones for the purposes of a problem. However, one must be careful, because a common mistake is to assume that all options are equally likely in a system.

Example 4.1. Suppose there are n balls to be placed in m urns (e.g. strings in a hash table, server requests to m machines in a computer cluster, etc.). The counts of the balls in the urns aren't equally likely. For example, if there are 2 balls A and B and two urns 1 and 2, then all possibilities are equally likely, since each ball has a $1/2$ probability of ending up in each particular urn.

But once A and B are taken to be indistinguishable, then the counts differ: there is one configuration in which both balls end up in urn 1, but two in which one ball ends up in each urn. Be careful!

Now, consider two fair, six-sided dice, yielding values D_1 and D_2. Let E be the event $D_1 + D_2 = 4$. Then, $P(E) = 1/12$, since there are 3 possibilities out of 36 total. Then, let F be the probability that $D_1 = 2$. What is $P(E)$ if F has already been observed? Not all 36 rolls are possible, and S is reduced to $\{(2,1),(2,2),\ldots,(2,6)\}$, and the event space is $\{(2,2)\}$. Thus, the probability $P(E)$ given F is $1/6$ — which is different. This basic idea is one of the most important concepts of probability.

Definition. A conditional probability is the probability of an event E occurs given that some other event F has already occured. This is written $P(E \mid F)$.

In this case, the sample space is reduced to those options consistent with F, or $S \cap F$, and the event space is reduced in the same way to $E \cap F$. Thus, in the case of equally likely outcomes, $P(E \mid F) = |E \cap F|/|S \cap F| = |E \cap F|/|F|$, because $F \subseteq S$.

This can be generalized to the general case, even when there aren't equally likely outcomes: $P(E \mid F) = P(E \cap F)/P(F)$, where $P(F) > 0$. Probabilities are used here because counts aren't as valid. This implies something known as the "chain rule," which shows that $P(E \cap F) = P(E \mid F)P(F)$. Intuitively, the probability of both E and F happening is the probability that F happens, multiplied by the probability that one happens, then the other happens given the first. This also happens to be commutative, so this is also equal to $P(F \mid E)P(E)$. If $P(F) = 0$, then the conditional probability is undefined, because the statement "$P(E)$ given that F happened" makes no sense when F is impossible.

The chain rule is also known as the multiplication rule. The generalized version is

$$P\left(\bigcap_{i=1}^{n} E_i\right) = \prod_{i=1}^{n} P\left(E_i \,\middle|\, \bigcap_{j=1}^{i-1} E_j\right) = P(E_1)P(E_2 \mid E_1)P(E_3 \mid E_1 \cap E_2)\cdots P(E_n \mid E_1 \cap \cdots \cap E_{n-1}).$$

Note that this doesn't require the events to be mutually exclusive, which makes it quite powerful.

Example 4.2. Suppose 24 emails are sent to four users, and are evenly distributed. If 10 of the emails are spam, what is the probability user 1 receives 3 pieces of spam and user 2 receives 6 pieces of spam? Call the event for user 1 E and for user 2 F. Then,

$$P(E \mid F) = \frac{\binom{4}{3}\binom{14}{3}}{\binom{18}{6}} \approx 0.0784.$$

Here, user 2 can be thought of as a honeypot, which makes it less likely that spam reaches a legitimate user assuming the spammer is rate-limited.

If the above also has G, which is the event that user 3 receives 5 spam emails, $P(G \mid F) = 0$, since there aren't that many spam emails.

Example 4.3. Suppose a bit string with m 0s and n 1s is sent over a network, so all arrangements are equally likely. Then, if E is the probability that the first bit received is a 1, and F is the probability that k of the first r bits is a 1, then

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{P(F \mid E)P(E)}{P(F)}$$
$$P(F \mid E) = \frac{\binom{n-1}{k-1}\binom{m}{r-k}}{\binom{m+n-1}{r-1}}.$$

Notice that the bits are treated as distinct objects, which makes life easier. Then, $P(E) = n/(m+n)$, which is just counting how many of the bits are there, and $P(F) = \binom{n}{k}\binom{m}{r-k}/\binom{m+n}{r}$. Thankfully, things cancel out, and in the end $P(E \mid F) = k/r$.

Another way to think of this is that once we know there are k bits out of the first r are 1, then that's all that needs to be worried about: given this much smaller set, there are k possibilities for E among r choices given F.

Example 4.4. Suppose a deck of 52 cards is distributed into four piles, with 13 cards per pile. What is the probability that each pile contains exactly one ace? Let

- E_1 be the probability that the ace of spades is in any one pile,
- E_2 be the probability that the ace of spades and ace of hearts are in different piles,
- E_3 be the probability that the aces of spades, hearts, and clubs are in different piles, and
- E_4 the solution: that every ace is in a different pile.

Then, $P(E_1) = 1$, and $P(E_2 \mid E_1) = 39/51$, since there are 39 cards not in the pile with the ace of spaces. Then, there are 26 cards left in the other piles, so $P(E_3 \mid E_1 \cap E_2) = 26/50$, and by the same logic $P(E_4 \mid E_1 \cap E_2 \cap E_3) = 13/49$. Thus, the overall probability is

$$P(E_4) = P(E_1 \cap E_2 \cap E_3 \cap E_4) = \frac{39 \cdot 26 \cdot 13}{51 \cdot 50 \cdot 49} \approx 0.105.$$

Notice that a seemingly-convoluted problem becomes much easier from this viewpoint.

One of the most influential figures in the history of probability was Thomas Bayes,[8] who formulated an extremely important theorem named after him. It has several fomulations: $P(F \mid E) = P(E \cap F)/P(E) = (P(E \mid F)P(F))/P(E)$. (Yes, it really is that simple.)

The textbook writes Bayes' theorem as

$$P(E) = P(E \cap F) + P(E \cap (\sim F)) = P(E \mid F)P(F) + P(E \mid \sim F)P(\sim F).$$

These can be combined into the most computationally useful one:

$$P(F \mid E) = \frac{P(E \mid F)P(F)}{P(E \mid F)P(F) + P(E \mid \sim F)P(\sim F)}.$$

In the fully general form:

Theorem 4.1. *Let $F_1, \ldots, F_n$ be a set of mutually exclusive and exhaustive events (i.e. the sample space is the union of the F_i). Then,*

$$P(F_j \mid E) - \frac{P(E \cap F_j)}{P(E)} = \frac{P(E \mid F_j)P(F_j)}{\sum_{i=1}^{n} P(E \mid F_i)P(F_i)}.$$

This is useful if given that an event E has occured, one wishes to know whether one of the F_j also occured.

Example 4.5. Consider a test for HIV, which is 98% effective and has a false positive rate of 1%. It is known that about 1 person in 200 has HIV in the United States, so let E be the probability that someone tests positive for HIV with this test, and F be the probability that person actually has HIV. Then, $P(E \mid F)$ is the efficacy of the test, and the probability that a positive result is true is

$$\begin{aligned} P(F \mid E) &= \frac{P(E \mid F)P(F)}{P(E \mid F)P(F) + P(E \mid \sim F)P(\sim F)} \\ &= \frac{(0,98)(0,005)}{(0,98)(0.005) + (0.01)(1 - 0.005)} \approx 0.330. \end{aligned}$$

[8] ...who according to Dr. Sahami looks very much like Charlie Sheen. I don't see it.

Oddly enough, that the test that has such a high effectiveness still doesn't mean it returns the expected results! This is because only a very small number of people actually have HIV relative to the false positive rate, so a false positive is more likely than a real positive.[9] Consider the four possibilities: Table 1 contains a lot of information, and can be used

TABLE 1.

	HIV+	HIV-
Test positive	$0.98 = P(E \mid F)$	$0.01 = P(E \mid \sim F)$
Test negative	$0.02 = P(\sim E \mid F)$	$0.99 = P(\sim E \mid \sim F)$

to aid in calculations using Bayes' Theorem.

Example 4.6. Turning again to spam detection, suppose 60% of all email is spam, and 90% of spam has a forged header, but 20% of non-spam has a forged header. One can check for a forged header, and let E be the event that a message has a forged header, and F be the event that it is spam. Then, using Bayes' Theorem, $P(F \mid E) = (0.9)(0.6)/((0.9)(0.6) + (0.2)(0.4)) \approx 0.871$, so this allows for a really good characterization of whether email is spam.

Example 4.7. Consider the Monty Hall problem, in which there are 3 doors A, B, and C. Behind one door is a prize, and behind the other two is absolutely nothing. A contestant chooses a door, and then the game show host opens one of the other doors to show it is empty. Then, the contestant can switch the door it chooses, but is this a good idea?

- If the contestant doesn't switch, then the probability of winning is $1/3$, since opening the door doesn't make a difference.
- If it does switch, then without loss of generality assume the contestant chose A. If A was the winner, then the probability of winning after switching is zero. If one of the other two cases occurs (each of which happens with equal probability $1/3$), the contestant wins. Thus, the overall probability of winning is $2/3$, so it is advantageous to switch.

This is generally confusing or hard to understand, so tread carefully. Suppose instead of three doors there were 1000. Then, the host opens 998 empty doors. What is the chance the winner is in the remaining door? Awfully high, since the chance that you picked correctly initially is very small.

5. More Conditional Probability: 4/10/13

> "A bad statistician, afraid of terrorists when flying, brought a bomb on every flight, under the reasoning that it was extremely unlikely that there would be *two* bombs on a flight!"[10]

Definition. The odds of an event A is defined as $P(A)/P(\sim A) = P(A)/(1 - P(A))$. This is a positive value, but is not bounded above.

Thus, one can reason about conditional odds. Given some event E, the odds of some hypothesis H is

$$\frac{P(H \mid E)}{P(\sim H \mid E)} = \frac{P(H)P(E \mid H)/P(E)}{P(\sim H)P(\sim H \mid E)/P(E)} = \frac{P(H)P(E \mid H)}{P(\sim H)P(E \mid \sim H)}.$$

When E is observed, the odds of H are multiplied by $P(E \mid H)/P(E \mid \sim H)$, in some sense the odds that H is related to E being observed.

In Bayes' Theorem, one has $P(H \mid E) = P(E \mid H)P(H)/P(E)$. Here, $P(H)$ is called the prior, because it is the probability before any conditional expectation, $P(E \mid H)$ is called the likelihood, and $P(H \mid E)$ is called the posterior (after the condition). This is a form of probabilistic inference, which uses probability to determine whether a hypothesis is true after some evidence.

Example 5.1. An urn contains two coins A and B, where coin A comes up heads with probability $1/4$, and B returns heads with probability $3/4$. If a random coin was chosen from the urn and comes up heads when flipped, what are the odds that A was picked?

$$\frac{P(A \mid H)}{P(\sim A \mid H)} = \frac{P(A)P(H \mid A)}{P(\sim A)P(H \mid \sim A)} = \frac{P(A)P(H \mid A)}{P(B)P(H \mid B)} = \frac{(1/2)(1/4)}{(1/2)(3/4)} = \frac{1}{3}.$$

This is odds, not probability; be careful! It means that there is a $1/3$ ratio for A to B, so the probability becomes $1/4$.

A priori, before the coin was flipped, the odds of picking A was $1 : 1$, and the probability was $1/2$.

Here are some useful formulas for probability:

[9] Relevant xkcd.

[10] On the subject of abuses of conditional probability, here is a relevant xkcd.

- Commutativity: $P(A \cap B) = P(B \cap A)$.
- The chain rule (multiplication rule), as seen above: $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$.
- The intersection rule: $P(A \cap (\sim B)) = P(A) - P(A \cap B)$.
- Bonferroni's inequality: $P(A \cap B) \geq P(A) + P(B) - 1$. This will be proved in the next problem set, but it's a four-line proof if you do it right.

Another notion is the generality of conditional probability: for any events A, B, and E, these all still hold in the presence of event E:

- $P(A \cap B \mid E) = P(B \cap A \mid E)$, and $P(E \mid A \cap B) = P(E \mid B \cap A)$.
- The chain rule becomes $P(A \cap B \mid E) = P(A \mid B \cap E)P(B \mid E)$.
- Bayes' Theorem: $P(A \mid B \cap E) = P(B \mid A \cap E)P(A \mid E)/P(B \mid E)$. This is the same as before, but every probability has been conditioned on E.

There's actually some philosophy behind this, but this can be seen as E as a summary of everything you already know, and in some sense, every probability is a conditional probability, and without any such event E, there's no way to know any probability. How would you know the probabilities of a fair coin without any context? Formally, this means that $P(\quad \mid E)$ always satisfies the axioms of probability.

Example 5.2. For a simple example, take two six-sided dice D_1 and D_2. Let E be the probability that D_1 rolls a 1, and F be that D_2 rolls a 1. Then, $P(E) = P(F) = 1/6$, and $P(E \cap F) = 1/36$. Here, $P(E \cap F) = P(E)P(F)$.

Define a new event G, which is the probability that $D_1 + D_2 = 5$. Here, $P(G) = 1/9$, but $P(E \cap G) = 1/36 \neq P(E)P(G)$.

Definition. If E and F are two events such that $P(E \cap F) = P(E)P(F)$, then E and F are called independent events; otherwise, they are called dependent events.

Independent events can be thought of as events which aren't affected by each other. They also have the property (via the chain rule) that $P(E \mid F) = P(E)$. Observing F doesn't change the probability of E.

Claim. If E and F are independent events, then $P(E \mid F) = P(E \mid \sim F)$.

Proof.

$$\begin{aligned} P(E \cap (\sim F)) &= P(E) - P(E \cap F) \\ &= P(E) - P(E)P(F) \\ &= P(E)(1 - P(F)) \\ &= P(E)P(\sim F), \end{aligned}$$

so E and $\sim F$ are indepedent events, so $P(E \mid \sim F) = P(E) = P(E \mid F)$. The independence of E and F are necessary in the second line, so that $P(E \cap F) = P(E)P(F)$. ⊠

Definition. More generally, the events $E_1, \ldots, E_n$ are independent if for every subset $E_1, \ldots, E_r$ ($r \leq n$) of $E_1, \ldots, E_n$, it holds that

$$P\left(\bigcap_{i=1}^{r} E_1\right) = \prod_{i=1}^{r} P(E_i).$$

This just means that the product formula holds for all possible subsets. This sounds hairy, but isn't all that bad in practice.

Example 5.3. Roll two dice D_1 and D_2, and let E be the event that $D_1 = 1$, and F be that $D_2 = 6$. Then, let G be the event that $D_1 + D_2 = 7$. E and F are clearly independent. $P(E) = 1/6$ and $P(G) = 1/6$, but $P(E \cap G) = 1/36$, so they have to be independent. This is counterintuitive, but 7 is the exception, since there's always exactly one possibility for getting a 7 on D_2 no matter the roll on D_1. Thus, it is important to be careful when mathematics and intuition meet. Note also that F and G are independent by the same reasoning.

For more weirdness with intuiton, E and F are independent, E and G are independent, and F and G are independent, but E, F, and G are *not* independent: $P(E \cap F \cap G) = 1/36$, but $P(E)P(F)P(G) = (1/6)(1/6)(1/6) = 1/216$.

Example 5.4. Suppose a computer produces a sequence of random bits, where p is the probability that a bit is a `1`. Then, each bit is generated in an independent trial. Let E be the probability of getting n `1`s followed by a `0`.

The probability of n `1`s is p^n, and that of getting a `0` next is $1 - p$, so $P(E) = p^n(1 - p)$.

Example 5.5. Imagine a coin that comes up heads with probability p (really the same thing as the previous example). Then, the probability of n heads in n flips is p^n, and the probability of n tails in n flips is $(1 - p)^n$.

The probability of k heads first, then all tails is $p^k(1-p)^{n-k}$. This requires order, so if the order is ignored, then it's a matter of combination, so the probability is $\binom{n}{k}p^k(1-p)^{n-k}$. There are k heads to place in n slots if order is ignored. One could also place the $n-k$ tails rather than the heads, and this is the same, since $\binom{n}{k} = \binom{n}{n-k}$.

Example 5.6. Consider a hash table in which m strings are hashed into n buckets with equal, independent probability. Let E be the probability that at leat one string was hashed into the first bucket. In general, defining one's own events makes solving these problems much easier by breaking them into subproblems. Additionally, the phrase "at least one" suggests solving by taking the complement, which is that no strings are hashed into the first bucket.

Let F_i be the probability that string i is not hashed into the first bucket, so $P(F_i) = 1 - 1/n = (n-1)/n$ for every i. Then, $E = \sim\bigcap_{i=1}^{m} F_i$, so

$$P(E) = 1 - P\left(\bigcap_{i=1}^{m} F_i\right) = 1 - \prod_{i=1}^{m} P(F_i) = 1 - \left(\frac{n-1}{n}\right)^m.$$

This is very similar to the birthday problem, which attempts to find collisions in a calendar rather than a hash table.

Now, take m strings placed into a hash table with unequal probability: let p_i be the probability that a string is hashed into bucket i. This is a more accurate model for the real world, and has lots of applications. Then, let E be the probability that at least one of buckets 1 through k has at least one string hashed into it. Let F_i be the probability that at least one string is hashed into the i^{th} bucket; then, using De Morgan's Law,

$$P(E) = P\left(\bigcup_{i=1}^{k} F_i\right) = 1 - P\left(\sim\bigcup_{i=1}^{k} F_i\right) = 1 - P\left(\bigcap_{i=1}^{k} \sim F_i\right).$$

This is a problem, because these events aren't independent! If only one string is hashed, for example, knowing where it isn't adds some information about where else it might be. The strings are hashed independently, but the events F_i are about buckets, not strings.

It looks like we're stuck, but thinking about strings, the probability that it'll not be hashed into the first bucket is $1-p_1$, and the probability that it won't be in the first two is $1-p_1-p_2$. This illustrates the technique of stepping back and figuring out what it all means. Thus, the probability that no strings are hashed into buckets 1 to k is

$$P\left(\bigcap_{i=1}^{k} \sim F_i\right) = \left(1 - \sum_{i=1}^{k} p_i\right)^m,$$

so the probability of the original event is

$$P(E) = 1 - \left(1 - \sum_{i=1}^{k} p_i\right)^m.$$

Considering one final question about this hash table, let E be the probability that each of the buckets 1 to k has a string hashed into it. Then, if F_i is the probability that at least one string is hashed into the i^{th} bucket. These events aren't independent, so this becomes

$$\begin{aligned} P(E) &= P(F_1 \cap \cdots \cap F_k) = 1 - P(\sim(F_1 \cap F_2 \cap \cdots \cap F_k)) \\ &= 1 - P\left(\bigcup_{i=1}^{k} \sim F_i\right) \\ &= 1 - \sum_{r=1}^{k} (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(\sim(F_{i_1}) \cap \cdots \cap (\sim F_{i_r})), \end{aligned}$$

where $P(\sim F_{i_1} \cap \cdots \cap \sim F_{i_r}) = (1 - p_{i_1} - \cdots - p_{i_r})^r$ as before, using the union formula.

Before the next example recall the geometric series

$$\sum_{i=0}^{n} x^i = \frac{1-x^{n+1}}{1-x}.$$

As $n \to \infty$, if $|x| < 1$, then the infinite sum reduces to $1/(1-x)$ in the limit.

Example 5.7. Suppose two six-sided dice are rolled repeatedly. What is the probability that a 5 is rolled before a 7? This is a simplified version of a game called craps. Define the series of events F_n, in which neither a 5 nor a 7 is rolled

in the first $n-1$ trials, but a 5 was rolled on the n^{th} trial. Thus, the probability of rolling a 5 on any trial is $4/36$, and of a 7 on any trial is $6/36$. Thus, the overall probability is

$$P(E) = P\left(\bigcup_{i=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} P(F_n) = \sum_{n=1}^{\infty} \left(1 - \left(\frac{10}{36}\right)\right)^{n-1} \left(\frac{4}{36}\right) = \sum_{n=1}^{\infty} \left(\frac{26}{46}\right)^{n-1} \left(\frac{4}{36}\right) = \frac{4}{36} \sum_{n=0}^{\infty} \left(\frac{26}{36}\right)^{n} = \frac{2}{5}$$

after using the geometric series. Notice that the events F_i are all mutually exclusive, so the probability of their union is just the sum of their probabilities.

6. Conditional Independence and Random Variables: 4/12/13

"There are three kinds of lies: lies, damned lies, and statistics." – misattributed to Mark Twain

Returning to Example 5.7, there is a simpler way to look at it: the goal is to roll a 5 before rolling a 7, which is $P(5)/P(5 \text{ or } 7) = (4/36)/(10/36) = 2/5$. This is a much easier way to solve the problem, but it requires a little more thought.

Example 6.1. Consider DNA paternity testing, in which a child is born with a pair of genes (A, a), which will be event $B_{A,a}$. The mother has a gene pair (A, A), and there are two possible fathers, $M_1 = (a, a)$ and $M_2 = (a, A)$. Suppose the probability that M_1 is the father is p (ignoring the child's genes), so the probability that M_2 is the father is $1 - p$. Then,

$$P(M_1 \mid B_{A,a}) = \frac{P(M_1 \cap B_{A,a})}{P(B_{A,a})} = \frac{P(B_{A,a} \mid M_1)P(M_1)}{P(B_{A,a} \mid M_1)P(M_1) + P(B_{A,a} \mid M_2)P(M_2)} = \frac{1 \cdot p}{1 \cdot p + (1-p)/2} = \frac{2p}{1+p} > p.$$

Thus, this genetic test makes it more likely that M_1 is the father. *A priori*, it makes sense that M_1 is the father, since there are more ways for M_1 to contribute an a to the child than M_2 to.

One standard class of problems in probability is called coupon collecting: there exist N different sorts of coupons, and each is collected with some probability p_i. Then, one can ask questions such as:

- After collecting m coupons, what is the probability that there are k different kinds?
- What is the probability that one has at least one of each coupon after collecting m coupons?

These problems come up a lot in computer science, and have in fact come up before in this course, such as in Example 5.6.

Recall that the laws of probability work when conditioning was thrown in, but this doesn't hold true for indepedence (a property): it is *not* necessarily true that if E and F are independent events, that $P(E \mid G)P(F \mid G) = P(E \cap F \mid G)$. However, if all of the events are independent, then equality does hold.

Example 6.2. Roll two six-sided dice, giving values D_1 and D_2. Then, let E be the event $D_1 = 1$, F be the event $D_2 = 6$, and G be $D_1 + D_2 = 7$. E and F are independent, as was shown in Example 5.3. However, $P(E \mid G) = 1/6$, $P(F \mid G) = 1/6$, and $P(E \cap F \mid G) = 1/6$. Introducing new information can do weird things to independence.

Example 6.3. Suppose a dorm has 100 students, and 10 of them are CS majors. 30 students get straight As, and there are 3 CS majors who get straight As. Let C be the probability that a randomly chosen student is a CS major, and A be that a randomly chosen student gets straight As. Thus, $P(C \cap A) = 0.03 = P(C)P(A)$, so C and A are independent.

Now, it's faculty night, which means that only CS majors and A students show up, so 37 students arrive. 10 of them are CS majors, and 30 are A students, so $P(C \mid C \cup A) = 10/37$. It looks like being a CS major lowers the probability of getting good grades... which isn't true. Once Faculty Night, a new event, is introduced, they are dependent.

This can be explained away (no, really; this line of reasoning is called explaining away). Everyone at Faculty Night needs a reason for showing up. If a lawn is watered by either rain or sprinklers in independent events. If the grass is wet, and then someone notices that the sprinklers are on, does that make it less likely that it has rained? In general, if there are multiple things that can cause an observation and one of them is observed, that makes the other causes less likely, since the event is already accounted for.

Alternatively, if you go to the doctor with a cough, and the doctor says you might have a cold or dengue fever, and the doctor says that a test result indicates you have a cold, then do you think you have dengue fever?

Not only can conditioning create dependence, it can destroy it.

Example 6.4. Let A be the event that it is not Monday on a randomly chosen day of the week, let B be that it is Saturday, and let C be that it is the weekend.

Intuitively, A and B are dependent, and mathematically $P(A) = 6/7$, $P(B) = 1/7$ and $P(A \cap B) = 1/7$. However, after conditioning on C, $P(A \mid C) = 1$, $P(B \mid C) = 1/2$, and $P(A \cap B \mid C) = 1/2 = P(A \mid C)P(B \mid C)$. Set-theoretically, this depends on how exactly the events are dependent. A field of research called Bayesian networks investigates this; take CS 228 for more details.

Definition. Two events E and F are called conditionally independent given an event G if $P(E \cap F \mid G) = P(E \mid G)P(F \mid G)$, or, equivalently, $P(E \mid F \cap G) = P(E \mid G)$.

Exploiting conditional independence to make probabilistic computations faster was a major application of computer science on probability theory (as opposed to the other way around).

Definition. A random variable is a real-valued function on a sample space.

This sounds kind of unclear, but one simple example is the number of heads obtained when three coins are flipped. Intuitively, a random variable is something that can be measured as a result of observing some probabilistic experiment. If Y is the random variable indicating the number of heads, then $P(Y = 0) = 1/8$, $P(Y = 2) = 3/8$, and $P(Y = 5) = 0$.

Definition. A binary random variable is a random variable with two possible values.

This is the simplest possible random variable (other than the constant one, which is silly). Flipping a coin is a good example of this, but keep in mind the results aren't H and T, but real values corresponding to heads and tails.

Example 6.5. Consider an urn with 11 balls: 3 are red, 3 are blue, and 5 are black. Then, suppose one draws 3 balls, and is awarded \$1 for a blue ball, $-$\$1 for a red ball, and 0 for the black ones. Then,

$$P(Y = 0) = \frac{\binom{5}{3} + \binom{3}{1}\binom{3}{1}\binom{5}{1}}{\binom{11}{3}} = \frac{55}{165},$$

by considering all of the ways it can happen. Similarly,

$$P(Y = 1) = P(Y = -1) = \frac{\binom{3}{1}\binom{5}{2} + \binom{3}{2}\binom{3}{1}}{\binom{11}{3}} = \frac{39}{165}.$$

$P(Y = k) = P(Y = -k)$ by symmetry.

This can be generalized:

Definition. A random variable is discrete if it can take on countably many values.

Definition. A probability mass function p (PMF) of a discrete random variable X is $p(a) = P(X = a)$.

Thus, if X can take on values $x_1, x_2, \ldots$, then $\sum_{i=1}^{\infty} p(x_i) = 1$, so in general

$$P(X = a) = \begin{cases} p(x_i) \geq 0, & a = x_i \\ 0, & a \neq x_i \text{ for any } i \end{cases}$$

Another way to understand a PMF is as a histogram of the potential values.

Definition. For a random variable X, the Cumulative Distribution Function (CDF) is defined as $F(a) = P(X \leq a)$, with $a \in \mathbb{R}$.

This isn't terribly helpful in the discrete case, but will be much more essential in the continuous case, which will pop up in a couple weeks. That said, the CDF of a discrete random variable is $F(a) = \sum_{x \leq a} p(x)$.

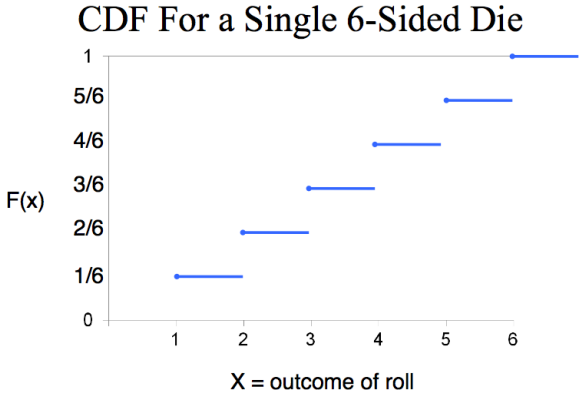


FIGURE 1. The CDF for a fair six-sided die. Source

Definition. The expected value for a discrete random value X is

$$E[x] = \sum_{\{x|p(x)>0\}} xp(x).$$

It's necessary to specify that $p(x) > 0$ for various tricky analytic reasons. This is also known as the mean, the center of mass, the average value, etc.

Example 6.6. The expected value of rolling a six-sided die is $E[X] = \sum_{i=1}^{6} i(1/6) = 7/2$. Notice that you can't actually roll a $7/2$, but this is the average over a large number of rolls.

Expected value is one of the most useful notions in probability.

Definition. A variable I is called an indicator variable for an event A if its value is 1 if A occurs, and 0 if $\sim A$ occurs.

Then, $P(I = 1) = P(A)$, so $E[I] = (1)(P(A)) + (0)(P(\sim A)) = P(A)$. This seems pretty tautological, but it's a huge concept: the expected value of the indicator is the probability of the underlying event.

Now, we have the mechanics to lie with statistics. Suppose a class has 3 classes with 5, 10, and 150 students. If a class is randomly chosen with equal probability, then let X be the number of students in the class. Then, $E[X] = 5(1/3) + 10(1/3) + 150(1/3) = 55$. The average value is 55, which seems reasonable.

But why do so many classes seem so large? If a student is randomly chosen and Y is the number of people in that class, then $E[Y] = 5(5/165) + 10(10/165) + 150(150/165) \approx 137$, which is about twice as much! $E[Y]$ is the student's perception, and $E[X]$ is what is usually reported.

This example illustrates that a lot of statistics involves interpretation, so the upshot is to be careful when reading statistics.

If g is some real-valued function, then let $Y = g(X)$. Then, the expectation is

$$E[g(X)] = E[Y] = \sum_j y_j p(y_j) = \sum_j y_j \sum_{x_i : g(x_i) = y_i} p(x_i).$$

What this means is that one can obtain the probability for $Y = k$ by looking at the cases of X that lead to $Y = k$.

One consequence of this is linearity, in which g is a linear function. This means that $E[aX + b] = aE[X] + b$. For example, if X is the value of a 6-sided die and $Y = 2X - 1$, then $E[X] = 7/2$, so $E[Y] = 6$.

Definition. The n^{th} moment of a random variable X is

$$E[X^n] = \sum_{x:p(x)>0} x^n p(x).$$

This will be useful later; remember it!

The last concept for today is utility. If one has two choices with various consequences $c_1, \ldots, c_n$, where one of the c_i happens with probability p_i, and has a utility (cost) $U(c_i)$. For example. the probability of winning \$1 million in a \$1 lottery ticket is 10^{-7}. The utility of winning is \$99999 (since you have to buy the ticket), and of winning is -1. If the ticket isn't brought, you can't win, so the utility is zero. Then, the expected value of buying a ticket can be calculated to be $E \approx -0.9$. The takeaway lesson is that you can't lose if you don't play.

7. To Vegas: 4/15/13

"About the Binomial Theorem I am teeming with a lot o' news..." – The Major-General

There's an interesting puzzle of probability called the St. Petersburg Paradox. Suppose one has a fair coin (i.e. it comes up heads exactly half of the time). Let n be the number of coin flips before the first tails, and the winning value is $\$2^n$.

How much would you pay to play? In expectation, the payoff is infinite over a large number of trials, but that seems like a little much to pay. Formally, if X is the amount of winnings as a random variable,

$$E[X] = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{i+1} 2^i = \sum_{i=0}^{\infty} \frac{1}{2},$$

which is infinite. This seems weird, but mathematically, this is perfectly fine. Of course, for playing exactly once one wouldn't want to pay a million dollars...

Okay, now on to Vegas. Take some even-money game (e.g. betting red on a roulette table). Here, there is a probability $p = 18/38$ that one wins $\$Y$, and $(1 - p)$ that you lose $\$Y$. Make bets according to the following algorithm:

(1) Let $Y = \$1$.
(2) Bet Y.
(3) If this bet wins, then halt. Otherwise, multiply Y by 2 and go to step 2

Let Z be the amount of winnings upon stopping. Then,

$$E[Z] = \sum_{i=0}^{\infty} \left(\frac{20}{38}\right)^i \left(\frac{18}{38}\right) \left(2^i - \sum_{j=0}^{i-1} 2^j\right) = \left(\frac{18}{38}\right) \sum_{i=0}^{\infty} \left(\frac{20}{38}\right)^i (1) = 1$$

using the geometric series formula. The expected value of the game is positive, which is interesting. Thus, it can be used to generate infinite money, right?

Well, maybe not:

- Real games have maximum betting amounts to prevent strategies such as this one (called Martingale strategies).
- You have finite money, and are psychologically disinclined to spend it.
- Casinos like kicking people out.

Thus, in practice, in order to win infinite money, you need to have infinite money.

There are probability distributions with the same expected value, but in which the probabilities are "spread out" over a larger area (e.g. a uniform distribution from 0 to 10 versus a large spike at 5). Thus, some techniques have been introduced:

Definition. If X is a random variable with mean μ, then the variance of X, denoted $\operatorname{Var}(X)$, is $\operatorname{Var}(X) = E[(X-\mu)^2]$.

Notice that the variance is always nonnnegative. The variance is also known as the second central moment, or the square of the standard deviation.

In practice, the variance isn't computed by that formula:

$$\begin{aligned}
\operatorname{Var}(X) &= E[(X-\mu)^2] \\
&= \sum_x (x-\mu)^2 p(x) \\
&= \sum_x (x^2 - 2\mu x + mu^2) p(x) \\
&= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - E[X]^2, \qquad (1)
\end{aligned}$$

since $E[X] = \mu$. This last formula is easier to use, and the quantity $E[X^2]$ in particular is called the second moment (different from the second *central* moment).

Variance is preserved under linear transformation: $\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X)$.

Proof.

$$\begin{aligned}
\operatorname{Var}(aX+b) &= E[(aX+b)^2] - (E[aX+b])^2 \\
&= E[a^2X^2 + 2abX + b^2] - (aE[X]+b)^2 \\
&= a^2E[X^2] + 2abE[X] + b^2 - (a^2(E[X])^2 + 2abE[X] + b^2) \\
&= a^2E[X^2] - a^2E[X]^2 = a^2 \operatorname{Var}(X).
\end{aligned}$$

This depends on noticing that the expected value of a constant is just the constant itself, which makes sense. ⊠

Definition. The standard deviation of X is $\operatorname{SD}(X) = \sqrt{\operatorname{Var}(X)}$.

This is convenient because it has the same units as X does, and is therefore easier to understand.

Definition. The Bernoulli random variable is an experiment that can result in success (defined as 1) or failure (defined as 0). The probability of success is denoted p, and of failure $1-p$. The phrase "X is a Bernoulli random variable with probability p of success" is shortened to $X \sim \operatorname{Ber}(p)$.

The expectation of such a variable is $E[X] = p$, since the expectation of an indicator variable is just its probability, and $\operatorname{Var}(X) = p(1-p)$, which is a little harder to show.

Definition. A binomial random variable X denotes the number of successes in n independent trials of some Bernoulli random variable $\operatorname{Ber}(p)$. Then, one writes $X \sim \operatorname{Bin}(n,p)$, and $P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}$.

By the Binomial Theorem, as $n \to \infty$, the probability eventually becomes 1. For example, if one flips three fair coins and takes X to be the number of heads, then $X \sim \operatorname{Bin}(3, 0.5)$.

Example 7.1. This has a direct application to error-correcting codes (also called Hamming codes) for sending messages over unreliable networks. Take some 4-bit message, and append three "parity" bits. The goal is to take the message and choose parity bits in a way such that there are three sets with four bits, and each set has an even number of 1s.

Then, if some bit is lost in transmission, it can be detected, and it can even be fixed, by taking all the intersection of the sets with odd numbers of 1s and the complements of those with an even number (i.e. the correct ones).

Suppose each bit in the example is flipped with probability 0.1 (which is much more than in the real world), and if X is the number of bits corrupted, then $X \sim \text{Bin}(7, 0.1)$. Thus, the probability that a correct message is received is $P(X = 0) + P(X = 1)$. This can be calculated to be 0.8503, but without the error-correcting codes, $X \sim \text{Bin}(4, 0.1)$ and $P(X = 0) \approx 0.6561$. This is interesting because the reliability is significantly improved despite only being a software update.

Example 7.2. Consider a setup in which a person has two genes for eye color, and brown is dominant over blue (i.e. if the child has a brown gene, then it has brown eyes, and if it has two blue genes, then it has blue eyes).

Suppose each parent has one brown and one blue eye. Then, what is the probability that 3 children out of four have brown eyes? The probability that a single child has blue eyes is $1/4$, and that it has brown eyes is thus $3/4$.

Each child is an independent trial (in more ways that one), so $X \sim (4, 0.75)$, so $P(X = 3) = \binom{4}{3}(0.75)^3(0.25) \approx 0.4219$. This is considerably easier than modeling all of the combinatioons.

Here are some properties of $X \sim \text{Bin}(n, p)$: consider the k^{th} moment

$$E[X^k] = \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

Then, one can use $i\binom{n}{i} = n\binom{n-1}{i-1}$ (which can be verified by chasing factorials in the expansions of the formulas), yielding

$$\begin{aligned} E[X^k] &= np \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} \\ &= np \sum_{j=0}^{n-1} (j-1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-(j+1)}, \quad \text{where } j = i-1 \\ &= npE[(Y+1)^{k-1}], \end{aligned}$$

where $Y \sim \text{Bin}(n-1, p)$. Then, setting $k = 1$, $E[X] = np$, and if $k = 2$, then $E[X^2] = npE[Y+1] = np((n-1)p+1)$, so the variance is $\text{Var}(X) = E[X^2] - (np)^2 = np(1-p)$. After all this math, think about what these mean: the expected value is what intuition would suggest. This also offers a proof for the Bernoulli case, since $\text{Ber}(p) = \text{Bin}(1, p)$.

Example 7.3. This can be used to ask how powerful one's vote is. Is it better to live in a small state, where it makes it more likely that the vote will change the outcome in that state, or a larger one, where there is a larger outcome if the state does swing?

Adding some (not quite realistic) numbers: Suppose there are $a = 2n$ voters equally likely to vote for either candidate[11] and the voter in question will be the deciding $a + 1^{\text{st}}$ vote. The probability that there is a tie is

$$P(\text{tie}) = \binom{2n}{n} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n = \frac{(2n!)}{n!n!2^{2n}},$$

which simplifies to $1/\sqrt{n\pi}$ after Stirling's Approximation $n! \approx n^{n+1/2}e^{-n}\sqrt{2\pi}$. This is exceptionally accurate for $n \geq 70$, which is reasonable for this situation.

Then, the power of a tie is its probability multiplied by the number of electoral votes: since a is the size of the state, then this becomes $c\sqrt{2a/\pi}$, which would means that living in a larger state means a more powerful vote.

8. The Poisson Distribution: 4/17/13

"Life is good for only two things: doing mathematics, and teaching mathematics." – Simeon-Denis Poisson

It's possible to analyze the behavior of the binomial distribution in the limit: if $X \sim \text{Bin}(n, p)$, then $P(X = i) = n!(p^i(1-p)^{n-i})/(i!(n-i)!)$ Letting $\lambda = np$, this can be rewritten as

$$P(X = i) = \frac{n!}{i!(n-i)!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} = \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}.$$

If n is large, p is small, and λ is "moderate" (the exact meaning of which will be clarified in a second), then the first fraction is about 1, and $(1-\lambda/n)^n \approx e^{-\lambda}$ and $(1-\lambda/n)^i \approx 1$. In the limit, these all approach equality.

[11] so they're independent in at least two ways...

This can be used to approximate the binomial distribution for large n: $P(X = i) = \lambda^i e^{-\lambda}/i!$.

Definition. X is a Poisson[12] random variable, denoted $X \sim \text{Poi}(\lambda)$, if X takes on nonnegative integer values, and for the given parameter $\lambda > 0$,[13] it has a PMF distribution that $P(X = i) = e^{-\lambda}\lambda^i/i!$.

It's necessary to check that this is in fact a probability distribution: using a Taylor series, $e^\lambda = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$, so

$$\sum_{i=0}^{\infty} P(X = i) = \sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1.$$

Example 8.1. Suppose a string of length 10^4 is sent over a network and the probability of a single bit being corrupted is 10^{-6}. Using the Poisson distribution, $\lambda = (10^4)(10^{-6})$, $P(X = 0) = e^{-0.01}(0.01)^0/0! \approx 0.990049834$. Using a conventional binomial distribtuion, this is accurate to about nine decimal places, which is incredibly useful. See Figure 2 for a graph of the accuracy of the Poisson distribution as an approximation.

So what does "moderate" mean for λ, anyways? There are different interpretations, such as $n > 20$ and $p < 0.05$, or $n > 100$ and $p < 0.01$. Moderate really just means that the approximation is reasonably accurate.

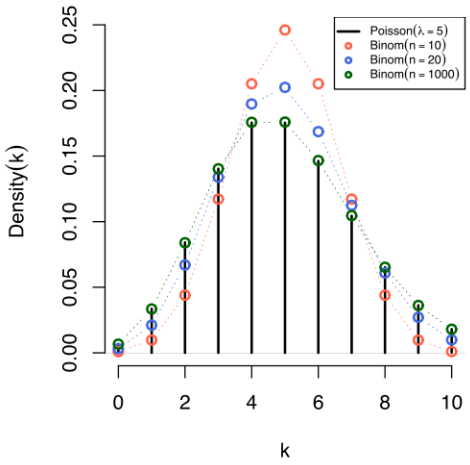


FIGURE 2. An illustration of the differences between binomial and Poisson distributions. Source

If $X \sim \text{Poi}(\lambda)$, where $\lambda = np$, then $E[X] = np = \lambda$, and $\text{Var}(X) = np(1-p) = \lambda(1-0) = \lambda$ as $n \to \infty$ and $p \to 0$. Thus, λ represents the number of trials that are expected to be true.

Example 8.2. Intuitively, imagine that you are baking an extremely large raisin cake, which is cut into slices of moderate size (with respect to the number of raisins in the slice of cake). The probability p that any particular raisin is in any particular slice of cake is vanishingly small; if there are n cake slices, $p = 1/n$, assuming an even distribution. If X is the number of raisins in a given cake slice, $X \sim \text{Poi}(\lambda)$, where $\lambda = R/n$, where R is the total number of raisins.

There are lots of concrete applications of this to computer science: one could sort strings into buckets in a hash table, determine the list of crashed machines in some data center, or Facebook login requests distributed across some set of servers.

Example 8.3. If computer chips are produced, so that $p = 0.1$ is the probability that a chip is defective. In a sample of $n = 10$ chips, what is the probability that a sample contains at most one defective chip? n and p aren't extreme, so using the binomial distribution, $P(Y \leq 1) \approx 0.7361$, and with the Poisson distribution, $P(Y \leq 1) \approx 0.7358$. This is still really close.

Computing $P(X \leq a)$ when $X \sim \text{Poi}(\lambda)$ straightforwardly is computationally expensive, but it turns out that $P(X = i+1)/P(X = i) = \frac{\lambda}{\lambda+1}$, so one can just do one approximation and then multiply, which is pretty fast.

A Poisson distribution is an approximation, so it still approximately works for things that aren't really binomial distributions. This is known as the Poisson paradigm. If the dependency is mild, then the approximation is still pretty close. For example, in a large hash table, the number of entries in a bucket are dependent events, but it grows weaker with the size of the hash table. Additionally, if the probability of success in each trial varies slightly, the Poisson

[12] Poisson literally means "fish" in French, though this is not the reason that this distribution has its name.

[13] Relevant xkcd.

distribution still works pretty well (e.g. when load on a network varies slightly with time). In each of these cases, the binomial random variable is also a good approximation, but it comes with the connotation that the probability is exact, whereas the Poisson distribution doesn't.

Returning to the birthday problem, take m people and let $E_{x,y}$ be the event that people x and y have the same birthday. Thus, $P(E_{x,y}) = 1/365 = p$, but not all of these events are independent (e.g. if Alice and Bob have the same birthday and Bob and Eve have the same birthday, do Alice and Eve have the same birthday?). Thus, this isn't a binomial distribution, so let $X \sim \text{Poi}(\lambda)$, where $\lambda = p\binom{m}{2} = m(m-1)/730$. Then,

$$P(X = 0) = e^{-\frac{m(m-1)}{730}} \frac{(m(m-1)/730)^0}{0!} = e^{-\frac{m(m-1)}{730}}.$$

Then, the smallest m for which this probability is less than 0.5 can be solved to $m \geq 23$, which we know to be correct from before.

A Poisson process is a rare event that occurs over some time (e.g. radioactive decay, earthquakes, hits to a web server on the small scale). If there is some time interval over which these events are rare, then the events arrive at λ events per that interval of time. Then, split the time into n subintervals, where n is large. Events occuring in subintervals are independent, and within any particular subinterval, the probability of the event happening is small. Thus, a Poisson distribution can be used.

Example 8.4. Suppose a server averages two hits per second. Let X be the number of hits the server receives in a second. If $X \geq 5$, then the server crashes. How often does this happen?

Suppose the sever can process at most one request per millisecond, so $n = 1000$. In a particular millisecond, the chance of getting a hit is $p = 2/1000 = 1/500$. Then, $\lambda = np = 2$, $X \sim \text{Poi}(2)$, and $P(X = 5)$ can be calculated.

Definition. X is a geometric random variable, denoted $X \sim \text{Geo}(p)$, if X is a random variable denoting the number of independent trials with probability p of success on each trial s.t. all but the last trial are failures, and the last is success. Then, $P(X = n) = (1-p)^{n-1}p$.

Then, $E[X] = 1/p$ and $\text{Var}(p) = (1-p)/p^2$.

Example 8.5.

- Flipping a fair coin until the first heads appears.
- Drawing balls from an urn that contains both white and black with replacement until a white ball is drawn.
- Generating random bits until a `1` is gnerated.

This seems a bit easier than the binomial distribution, so let's try another one:

Definition. X is a negative binomial random variable, denoted $X \sim \text{NegBin}(r, p)$, if X is the number of independent trials until r successes, where p is the probability of success on each trial. X takes on values $X \geq r$ with probability

$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r},$$

where $n \geq r$. Notice that there is no upper bound on n, though the probability decreases after some point.

The $\binom{n-1}{r-1}$ term isn't obvious, but the last trial must end in success, so the number of choices is restricted. Notice additionally that $\text{Geo}(p) \sim \text{NegBin}(1, p)$, so the geometric distribution is analogous to the Bernoulli distribution for the negative binomial distribution.

Definition. X is a hypergeometric random variable, denoted $X \sim \text{HypG}(n, N, m)$, if it varies as in the following experiment: consider an urn with N balls, in which $N - m$ are black and m are white. If n balls are drawn without replacement, then X is the number of white balls drawn.

This is just combinatorics: $P(X = i) = \binom{m}{i}\binom{N-m}{n-i}/\binom{N}{m}$, where $i \geq 0$. Then, $E[X] = n(m/N)$ and $\text{Var}(X) = (nm(N-n)(N-m))/(N^2(N-1))$.

Let $p = m/N$, the probability of drawing a white ball on the first draw. Note that as $N \to \infty$, $\text{HypG}(n, N, m) \to \text{Bin}(n, m/N)$. The binomial is the case with replacement, and the hypergeometric is the case without replacement. As mentioned above, one can be made to approximate the other.

Example 8.6. Suppose N is the number of individuals of some endangered species that remain. If m of them are tagged, and then allowed to mix randomly, then randomly observe another n of them. Let X be the number of those n that were tagged. Then, $X \sim \text{HypG}(n, N, m)$. Then, using something called a maximum likelihood estimate, $\hat{N} = mn/i$ is the value that maximizes $P(X = i)$.

9. From Discrete to Continuous: 4/19/13

The stuff we have been discussing has come up in the real world. The Supreme Court case *Berghuis v. Smith* dealt with determining how to detect whether a group is underrepresented in a jury pool.

> Justice Breyer opened the questioning by invoking the binomial theorem. He hypothesized a scenario involving "an urn w/ 1000 balls, 60 of which are red, 940 are black, and you select them randomly twelve at a time." He then estimated that if the red balls represented black jurors, then something like one-third to one-half of the juries would have a black person on them.

Justice Breyer was wrong, though, since the balls are drawn without replacement. Thus, this is a hypergeometric distribution, rather than binomial, and the probability that twelve black balls are drawn is $\binom{940}{12}/\binom{1000}{12} \approx 0.4739$, so the probability that at least one red ball is drawn is one minus that, or about 0.5261. Not quite what Justice Breyer got.

But this can be approximated with the binomial distribution, since in a large pool replacement doesn't make much of a difference. The probability doesn't change: the number of red balls drawn, X is $X \sim \text{Bin}(12, 60/1000)$, so $P(X \geq 1) = 1 - P(X = 0) \approx 1 - 0.4759 = 0.5240$, which isn't much different.[14] The upshot is that people make mistakes all the time in probability, and it's worth fact-checking. There's actually some nuances depending on how the jury selection is modeled, but the numbers don't change very much.

So far, the random variables seen in this class have been discrete. However, continuous random variables exist: these can take on an uncountable range of values, typically representing measurement (height, weight, time). The distinction is one between "how much?" and "how many?" Mathematically, this means replacing sums with integrals.

Definition. X is a continuous random variable if there is a nonnegative function $f(x)$ on $\mathbb{R}$ such that $P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$.

f is a probablity density function (PDF) if $\int_{-\infty}^{\infty} f(x) = 1$.

An important distinction to be made is that f is *not* a probability. The units don't match, and $f(x)$ can take on values greater than 1. There is the additional quirk that $P(X = a) = \int_a^a f(x)\,\mathrm{d}x = 0$. Continuous random variables are useful for ranges, not exact numbers. Contrast this with a probability mass function: in the countable case, individual values can have nonzero probabilities.

Definition. For a continuous random variable X, the cumulative distribution function (CDF) is $F(X) = P(X \leq a) = P(X < a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x$.

Thus, the density f is just the derivative of the CDF: $f(a) = \frac{\mathrm{d}F}{\mathrm{d}a}$. Then, if f is continuous and ε is small,

$$P\left(a - \frac{\varepsilon}{2} \leq X \leq a + \frac{\varepsilon}{2}\right) = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)\,\mathrm{d}x \approx \varepsilon f(a).$$

This is important, because it allows ratios of probabilities to be meaningfully defined:

$$\frac{P(X = 1)}{P(X = 2)} \approx \frac{\varepsilon f(1)}{\varepsilon f(2)} = \frac{f(1)}{f(2)},$$

since $\varepsilon \to 0$ is perfectly fine in the limit. Thus, things such as conditional probability are meaningful.

Example 9.1. Suppose X is a continuous random variable (CRV) with PDF

$$f(x) = \begin{cases} C(4x - 2x^2), & 0 < x < 2, \\ 0, & \text{otherwise.} \end{cases}$$

First, C can be calculated:

$$\int_0^2 C(4x - 2x^2)\,\mathrm{d}x = C\left[2x^2 - \frac{2x^3}{3}\right]_0^2 = 1,$$

so plugging in, $C((8 - 16/3) - 0) = 1$, so $C = 3/8$. Then,

$$P(X > 1) = \int_1^\infty f(x)\,\mathrm{d}x = \int_1^2 \frac{3}{8}(4x - 2x^2)\,\mathrm{d}x = \frac{3}{8}\left[2x^2 - \frac{2x^3}{3}\right]_1^2 = \frac{1}{2},$$

which is what makes sense (since the function is symmetric).

[14] The cause of misestimation may have been that CS 109 didn't exist when Justice Breyer was at Stanford...

Example 9.2. Suppose X is the amount of time (in days) before a disk crashes. It is given by the PDF

$$f(x) = \begin{cases} \lambda e^{-\frac{x}{100}}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

First, λ will be calculated:

$$1 = \int_0^\infty \lambda e^{-x/100}\,\mathrm{d}x = -100\lambda \int_0^\infty -\frac{e^{-x/100}}{100}\,\mathrm{d}x = -100\lambda e^{-x/100}\Big|_0^\infty = 100\lambda,$$

so $\lambda = 1/100$. Technically, an improper integral was taken here and sort of glossed over.

Then, what is $P(50 < X < 150)$?

$$F(150) - F(50) = \int_{50}^{100} \frac{1}{100} e^{-x/100}\,\mathrm{d}x = -e^{-x/100}\Big|_{50}^{150} \approx 0.383.$$

We can also calculate $P(X < 10)$:

$$F(10) = \int_0^{10} \frac{1}{100} e^{-x/100}\,\mathrm{d}x = -e^{-x/100}\Big|_0^{10} \approx 0095.$$

Expectation and variance can also be cacluated: if X is a continuous random variable,

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \qquad E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x \qquad E[X^n] = \int_{-\infty}^{\infty} x^n f(x)\,\mathrm{d}x.$$

Linearity of expectation is still a thing: $E[aX + b] = aE[X] + b$, and the same proof can be given. Similarly, the formulae for variance still hold.

Example 9.3. For example, if X has linearly increasing density, given by $f(x) = 2x$ on $[0, 1]$ and $f(x) = 0$ otherwise, the expectation isn't obvious from the graph, but can be found as

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x = \int_0^1 2x^2\,\mathrm{d}x = \frac{2}{3}$$

and the variance is given by

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = \int_0^1 2x^3\,\mathrm{d}x = \frac{1}{2},$$

so $\mathrm{Var}(X) = E[X^2] - E[X]^2 = 1/2 - (2/3)^2 = 1/18$.

These are arguably easier than in the discrete case.

Definition. X is a uniform random variable, denoted $X \sim \mathrm{Uni}(\alpha, \beta)$, if its probability density function is

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \alpha \leq x \leq \beta \\ 0, & \text{otherwise.} \end{cases}$$

This is just a horizontal line.

Sometimes, this is defined only on (α, β), but this causes issues later on. The probability is given by

$$P(a \leq x \leq b) = \int_a^b f(x)\,\mathrm{d}x = \frac{b - a}{\beta - \alpha},$$

which is valid when $\alpha \leq a \leq b \leq \beta$. The expectation is even easier: it can be formally calculated to be $(\alpha + \beta)/2$, which is no surprise, and the variance is (after a bunch of math) $\mathrm{Var}(X) = (\beta - \alpha)^2/12$.

Example 9.4. Suppose $X \sim \mathrm{Uni}(0, 20)$. Then. $f(x) = 1/20$ for $0 \leq x \leq 20$ and $f(x) = 0$ otherwise. Thus, $P(X < 6) = \int_0^6 \mathrm{d}x/20 = 6/20$, and $P(4 < X < 17) = \int_4^{17} \mathrm{d}x/20 = (17 - 4)/20 = 13/20$.

Example 9.5. Suppose the Marguerite bus arrives at 15-minute intervals (on the quarter hour), and someone arrives at a bus stop uniformly sometime between 2 and 2:30. Thus, $X \sim \mathrm{Uni}(0, 30)$. If the passenger waits less than 5 minutes for the bus, then it must arrive between 2:10 and 2:15, or 2:25 to 2:30. This is 10 minutes out of the 30, so the probability is $1/3$. (Since the distribution is uniform, this sort of calculation can be done.) The passenger waits for more than 14 minutes if it arrive between 2 and 2:01 or 2:15 to 2:16. These are 2 out of the 30 minutes, so the probability is $1/15$.

Example 9.6. Suppose a student bikes to class, and leaves t minutes before class starts, which is (usually) a choice by the student. If X is the travel time in minutes, given by a PDF of $f(x)$. Then, there is a cost to be early or late to class: $C(X,t) = c(t-X)$ if early, and $C(X,t) = k(t-X)$ if late (which is a different cost). Thus, the cost function is discontinuous. Then, choose t that minimizes $E[C(X,t)]$:

$$E[C(X,t)] = \int_0^\infty C(X,t)f(x)\,\mathrm{d}x = \int_0^t c(t-x)f(x)\,\mathrm{d}x + \int_t^\infty k(x-t)f(x)\,\mathrm{d}x.$$

Now we need to minimize it, which involves some more calculus. Specifically, it requires the Leibniz integral rule:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{f_1(t)}^{f_2(t)} g(x,t)\,\mathrm{d}x = \frac{\mathrm{d}f_2(t)}{\mathrm{d}t}g(f_2(t),t) - \frac{\mathrm{d}f_1(t)}{\mathrm{d}t}g(f_1(t),t) + \int_{f_1(t)}^{f_2(t)} \frac{\partial g(x,t)}{\partial t}\,\mathrm{d}x.$$

Let t^* be the optimal time. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E[C(X,t)] &= c(t-t)f(t) + \int_0^t cf(x)\,\mathrm{d}x - k(t-t)f(t)\int_t^\infty kf(x)\,\mathrm{d}x \\ 0 &= cF(t^*) - k(1-F(t^*)) \\ \implies F(t^*) &= \frac{k}{c+k}.\end{aligned}$$

The question comes down to whether being late is much worse than being early. If so, this becomes closer to 1, so t should be earlier. Similarly, if being early is a problem, then this becomes close to 0, so t should be later.

10. The Normal Distribution: 4/22/13

"Welcome to week four of CS 106A!"

Definition. X is a normal random variable (sometimes called Gaussian[15]), denoted $X \sim \mathrm{N}(\mu, \sigma^2)$, if it has a PDF of $f(x) = e^{-(x-\mu)^2/2\sigma^2}/(\sigma\sqrt{2\pi})$, where μ represents the mean and σ the standard deviation.

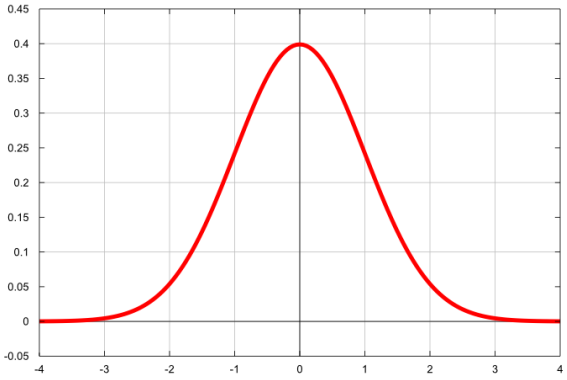


Figure 3. A normal distribution. Source

This takes on a nonzero value for all $x \in \mathbb{R}$. It has very nice moments: $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. This is a very common distribution for natural phenomena because it results from the sum of multiple roughly independent variables.

Here are some properties of a normal random variable:

- If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y = aX + b$, then by linearity of expectation $Y \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$, so $E[Y] = a\mu + b$ and $\mathrm{Var}(Y) = a^2\sigma^2$.
- Looking at the CDF, with X and Y as above,

$$F_Y(x) = P(Y \le X) = P(aX + B \le x) = P\left(X \le \frac{x-b}{a}\right) = F_X\left(\frac{x-b}{a}\right).$$

Obviously, this causes problems if $a = 0$, but that's a silly case. After differentiating, one obtains the PDF:

$$f_Y(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_Y(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X\left(\frac{x-b}{a}\right) = \frac{1}{a}f_X\left(\frac{x-b}{a}\right).$$

[15] Apparently C.F. Gauss, who was responsible for popularizing this distribution, looks like Martin Sheen.

- In the special case where $Z = (X - \mu)/\sigma$, this is just $a = 1/\sigma$ and $b = -\mu/\sigma$, so

$$Z \sim \mathsf{N}(a\mu + b, a^2\sigma^2) = \mathsf{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) = \mathsf{N}(0,1).$$

  This transformation will be useful.

A random variable $Z \sim \mathsf{N}(0,1)$ is called a standard (unit) normal random variable, and has $E[Z] = 0$, $\mathrm{Var}(Z) = 1$, and $\mathrm{SD}(Z) = 1$. However, the CDF of Z doesn't have a closed form, and must be solved numerically: denote the CDF of a unit normal $F_Z(z) = \Phi(z)$, which satisfies

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{e^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x}{\sigma\sqrt{2\pi}} = \int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x.$$

By symmetry, $\Phi(-z) = P(Z \le -z) = P(Z > z) = 1 - \Phi(z)$. Geometrically, flipping Φ about the y-axis converts the region $Z \le -z$ into $Z > z$, which has probability $1 - \Phi(z)$, and the transformation is area-preserving. This is important to understand because tables of the value of Φ don't store negative values, since they can be so easily calculated from the positive ones. Thus, Z can be used to compute the CDF for an arbitrary normal random variable $X \sim \mathsf{N}(\mu, \sigma^2)$, for $\sigma > 0$. Thus,

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right), \end{aligned}$$

with $Z \sim \mathsf{N}(a\mu + b, a^2\sigma^2)$. Then, one can use a table of values for $\Phi(z)$ to calculate the required answer.

Example 10.1. Suppose $X \sim \mathsf{N}(3, 16)$. Thus, $\mu = 3$ and $\sigma = 4$. Then, let Z be the unit normal, so that

$$\begin{aligned} P(X > 0) &= P\left(\frac{X-3}{4} > \frac{0-3}{4}\right) = P\left(Z > -\frac{3}{4}\right) = 1 - P\left(Z \le -\frac{3}{4}\right) \\ &= 1 - \Phi\left(-\frac{3}{4}\right) = 1 - \left(1 - \Phi\left(\frac{3}{4}\right)\right) = \Phi\left(\frac{3}{4}\right) \approx 0.7734 \end{aligned}$$

and

$$\begin{aligned} P(2 < X < 5) &= P\left(\frac{2-3}{4} < \frac{X-3}{4} < \frac{5-3}{4}\right) = P\left(-\frac{1}{4} < Z < \frac{2}{4}\right) \\ &= \Phi\left(\frac{2}{4}\right) - \Phi\left(-\frac{1}{4}\right) \approx 0.6915 - (1 - 0.5987) = 0.2902. \end{aligned}$$

More interestingly,

$$\begin{aligned} P(|X-3| > 6) &= P(X < -3) + P(X > 9) = P\left(Z < -\frac{3}{2}\right) + P\left(Z > \frac{3}{2}\right) \\ &= \Phi\left(-\frac{3}{2}\right) + 1 - \Phi\left(\frac{3}{2}\right) \approx 1(1 - 0.9332) = 0.1336. \end{aligned}$$

Other transformations might exist, but they can be solved in roughly the same way.

Example 10.2. Suppose a bit is sent on a wire with a voltage of either $2\,\mathrm{V}$ or $-2\,\mathrm{V}$ to represents bits of 1 and 0, respectively. Let X be the voltage sent and R be the voltage received. $R = X + Y$, where $Y \sim \mathsf{N}(0,1)$ is the error, called white noise. Then, R is decoded as 1 if $R \ge 0.5$, and 0 otherwise. What is the probability of an error in decoding given that $X = 1$? This is $P(2 + Y < 0.5) = P(Y < -1.5) = \Phi(-1.5) = 1 - \Phi(1.5) \approx 0.0668$, since Y is a standard normal random variable, making the computation a bit simpler. If $X = 0$, the probability of an error is instead $P(-2 + Y \ge 0.5) = P(Y \ge 2.5) = 1 - \Phi(2.5) \approx 0.0062$. These two are different because the decoding isn't symmetric involving the distribution of Y, which seems kind of odd, but happens all the time, in which the two cases are of different importance. For an analogy, suppose the bit represents a signal to launch the nuclear missles, or not; clearly, one kind of error is much worse than the other.

The normal approximation can be used to approximate the binomial distribution. Suppose $X \sim \mathrm{Bin}(n, p)$, so that $E[X] = np$ and $\mathrm{Var}(X) = np(1-p)$. The Poisson approximation is good for n large (about $n > 20$) and p small

(about 0.05 or less). For large n, we also have $X \approx Y \sim \mathsf{N}(E[X], \mathrm{Var}(X)) = \mathsf{N}(np, np(1-p))$, which is good when $\mathrm{Var}(X) \geq 10$. The deciding value is how spread the data is. Using this approximation,

$$\begin{aligned} P(X = k) &\approx P\left(k - \frac{1}{2} < Y < k + \frac{1}{2}\right) \\ &= \Phi\left(\frac{k - np + 0.5}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - np - 0.5}{\sqrt{np(1-p)}}\right). \end{aligned}$$

Here, something that is discrete is approximated by something continuous. This relies on the continuity correction, which integrates the continuous quantity over a step of size 1.[16]

Theorem 10.1 (DeMoivre-Laplace Limit Theorem)**.** *Let S_n be the number of successes in n independent trials, where success is given by probability p, then*

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b) - \Phi(a).$$

This theorem guarantees the approximation we have just taken. It provides the foundation for the Central Limit Theorem, which will be seen later, and illustrates why the normal distribution is so common in nature.

Example 10.3. Suppose 100 people are placed on a special diet, and X is the number of people on the diet whose cholesterol increases. Then, a doctor will endorse the diet of $X \geq 65$. What is the probability that the doctor endorses the diet despite it having no effect? This is represented by $X \sim \mathsf{Bin}(100, 0.5)$, so $np = 50$ and the standard deviation is 5. Then, we need the case $P(X \geq 65)$, where $\geq$ is distinct from $>$, because X is a discrete random variable.

Use a normal approximation: $P(X \geq 65) \approx P(Y > 64.5)$. (If it were necessary to calculate $P(X > 65)$, then one would use $P(Y > 65.5)$.) This is

$$P(Y \geq 64.5) = P\left(\frac{Y - 50}{5} > \frac{64.5 - 50}{5}\right) = 1 - \Phi(2.9) = 0.0019.$$

Here, the doctor is right.

Example 10.4. Suppose Stanford admits 2480 students and each student has a 68% chance of attending. If Stanford has room for 1745 students to matriculate, then let X be the number of students who attend, so that $X \sim \mathsf{Bin}(2480, 0.68)$. Thus, $\mu = 1686.4$ and $\sigma^2 \approx 539.65$, so this can be approximated by the normal variable $Y \sim \mathsf{N}(1686.4, 539.65)$, and

$$P(X > 1745) = P(Y \geq 1745.5) = P\left(\frac{Y - 1686.4}{23.23} > \frac{1745.5 - 1686.4}{23.23}\right) = 1 - \Phi(2.54) \approx 0.0055.$$

This is a good approximation, because $P(X > 1745) = 0.0053$, which is very close.

11. The Exponential Distribution: 4/24/13

> "Remember, don't drink and do probability. Not that I know from experience — and we're just talking about tea anyways."

Definition. X is an exponential random variable, denoted $X \sim \mathsf{Exp}(\lambda)$ for some $\lambda > 0$, if its PDF is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Then $E[X] = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$. This just looks like any decreasing exponential function, as in radioactive decay. Its CDF is $F(x) = 1 - e^{-\lambda x}$, since f is easy to integrate. This sort of distribution tends to represent the amount of time until some event (e.g. earthquakes, requests to a web server, etc.), which is why $x \geq 0$ and $\lambda > 0$.

This is a "memory-less" distribution, and in fact under reasonably mild conditions is the only such distribution. Let $X \sim \mathsf{Exp}(\lambda)$; then,

$$\begin{aligned} P(X > s + t \mid X > s) &= \frac{P(X > s + t \text{ and } X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)} \\ &= \frac{1 - F(s + t)}{1 - F(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} \\ &= 1 - F(t) = P(X > t). \end{aligned}$$

[16] One could use a different step size, but then some other constant for the step size would have to be multiplied in, which is unpleasant.

This means that the starting point is irrelevant: if you're waiting for something at time t with this distribution and s years have passed, then the expected value of waiting hasn't changed! In some sense, this means the curve is shaped like itself (which follows from the fact that $f' = \lambda f$).

Example 11.1. Suppose a visitor to some website leaves after X minutes, and on average a visitor leaves after 5 minutes. Thus, $X \sim \text{Exp}(1/5)$, since $E[X] = 5 = 1/\lambda$. Then, $P(X > 10) = 1 - F(10) = 1 - (1 - e^{-10\lambda}) = e^{-2} \approx 0.1353$ and $P(10 < X < 20) = F(20) - F(10) = (1 - e^{-4}) - (1 - e^{-2}) \approx 0.1170$.

Example 11.2. Let X be the number of hours of use until a laptop dies. On average, this is 5000, so $\lambda = 1/5000$. Then, make the rather conservative assumption that the laptop is used five hours per day. The probability that the laptop lasts four years is (ignoring leap years) $P(X > (5)(365)(4)) = P(X > 7300) = 1 - F(7300) = 1 - (1 - e^{-7300/5000}) = e^{-1.46} \approx 0.2322$. After five years, this becomes 0.1612 and after six, it becomes 0.1119. And thanks to the memoryless property, any one crash doesn't make the rest any less likely.

Recall some concepts from calculus: the product rule and integration by parts, which are just inverses of each other. This is useful for computing the n^{th} moments: $E[X^n] = x^n \lambda e^{-\lambda x} \, \mathrm{d}x$. Then, let $u = x^n$, so that $\mathrm{d}u = nx^{n-1} \, \mathrm{d}x$, and $v = -e^{-\lambda x}$, so $\mathrm{d}v = \lambda e^{-\lambda x} \, \mathrm{d}x$. Thus:

$$\begin{aligned} E[X^n] &= \int_0^\infty x^n \lambda e^{-\lambda x} \, \mathrm{d}x \\ &= \int_0^\infty u \, \mathrm{d}v = uv|_0^\infty - \int_0^\infty v \, \mathrm{d}u \\ &= -x^n e^{-\lambda x}\big|_0^\infty + \int_0^\infty n x^{n-1} e^{-\lambda x} \, \mathrm{d}x \\ &= \frac{n}{\lambda} \int_0^\infty x^{n-1} \lambda e^{-\lambda x} \, \mathrm{d}x = \frac{n}{\lambda} E[X^{n-1}]. \end{aligned}$$

This creates a recurrence that can be solved to generate $E[X^n] = n!/\lambda^n$. The first two moments are the most useful.

Definition. For two discrete random variables X and Y, define the joint probability mass function of X and Y to be $p_{X,Y}(a, b) = P(X = a, Y = b)$. Then, the marginal distribution of X is $p_X(a) = P(X = a) = \sum_y p_{X,Y}(a, y)$, and the marginal distribution of Y can be defined in the same way.

Example 11.3. Suppose a household has C computers, of which X are Macs and Y are PCs. Assume any given computer is equally likely to be a Mac or a PC. Then, define some probabilities: $P(C = 0) = 0.16$, $P(C = 1) = 0.24$, $P(C = 2) = 0.28$, and $P(C = 3) = 0.32$. Thus, one can make a joint distribution table by taking all the options that correspond to $C = c$ and splitting the probability given that each computer is equally likely to be of either type. Then, summing along the rows and columns gives the marginal distributions.[17]

This can be generalized to continuous distributions:

Definition. For two continuous random variables X and Y, their joint cumulative probability distribution is $F_{X,Y}(a, b) = F(a, b) = P(X \le a, Y \le b)$ where $a, b \in \mathbb{R}$.

Then, the marginal distributions are $F_X(a) = P(X \le a) = F_{X,Y}(a, \infty)$, and F_Y is defined in the same way.

This is no longer a table, but rather a two-dimensional graph, as in Figure 4.

Definition. Two random variables X and Y are jointly continuous if there exists a PDF $f_{X,Y}(x, y) : \mathbb{R}^2 \to \mathbb{R}$ such that

$$P(a_1 \le X \le a_2, b_1 \le Y \le b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x, y) \, \mathrm{d}y \, \mathrm{d}x.$$

and the CDF is given by

$$F_{X,Y}(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) \, \mathrm{d}y \, \mathrm{d}x.$$

Conversely, one can obtain the PDF from the CDF as $f_{X,Y}(a, b) = \frac{\partial^2 F_{X,Y}(a,b)}{\partial x \partial y}$. The marginal density functions are thus

$$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a, y) \, \mathrm{d}y \qquad \text{and} \qquad f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x, b) \, \mathrm{d}x.$$

[17] This is the source of the name: they're literally written in the margins of the table.

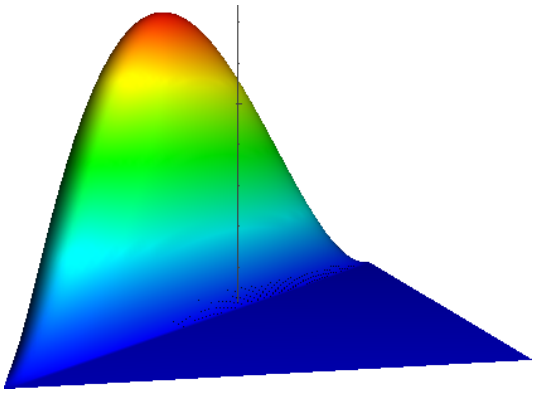


FIGURE 4. A joint probability distribution given by $F_{X,Y}(x,y) = 20(y-x)(1-y)$ for $x, y \geq 0$.

If these sorts of multiple integrals are foreign, see them as iterated integrals: integrate the innermost integral as in single-variable integrals, and treat all other variables as constants. This can be repeated, leading to a solution. For example,

$$\int_0^2 \int_0^1 xy \,\mathrm{d}x \,\mathrm{d}y = \int_0^2 \left(\int_0^1 xy \,\mathrm{d}x \right) \mathrm{d}y = \int_0^2 y \left[\frac{x^2}{2} \right]_0^1 \mathrm{d}y = \int_0^2 \frac{y}{2} \,\mathrm{d}y = \left[\frac{y^2}{2} \right]_0^2 = 1.$$

This can be generalized to three or more integrals, but that's not going to be required in this class. Intuitively, double integrals should be thought of as integrating over some area.

Additionally,

$$\begin{aligned} P(X > a, Y > b) &= 1 - P(\sim(X > a, Y > b)) = 1 - P((X \leq a) \cup (Y \leq b)) \\ &= 1 - (P(X \leq a) + P(Y \leq B) - P(X \leq A, Y \leq B)) \\ &= 1 - F_X(a) - F_y(B) + F_{X,Y}(a,b). \end{aligned}$$

Similary, $P(a_1 \leq X \leq a_2, b_1 \leq Y \leq b_2) = F_{X,Y}(a_1,b_1) + F_{x,y}(a_2,b_2) - F_{X,Y}(a_1,b_2) - F_{X,Y}(a_2,b_1)$. This can be seen by drawing a picture.

If X is a discrete random variable, then $E[X] = \sum_i iP(X=i)$, which can be generalized to the continuous case:

Lemma 11.1. *If Y is a non-negative continuous random variable with PDF $F(y)$, then*

$$E[Y] = \int_0^\infty P(Y > y) \,\mathrm{d}y = \int_0^\infty (1 - F(y)) \,\mathrm{d}y.$$

Proof.

$$\begin{aligned} \int_0^\infty P(Y > y) \,\mathrm{d}y &= \int_{y=0}^\infty \int_{x=y}^\infty f_Y(x) \,\mathrm{d}x \,\mathrm{d}y \\ &= \int_{x=0}^\infty \left(\int_{y=0}^\infty \mathrm{d}y \right) f_Y(x) \,\mathrm{d}x \\ &= \int_0^\infty x f_Y(x) \,\mathrm{d}x = E[Y]. \end{aligned}$$

⊠

12. The Multinomial Distribution and Independent Variables: 4/26/13

"So the moral of the story is to not be so *variable* in your arrival time."

The multinomial distribution is used to model n independent trials of some experiment, but each trial has one of m outputs (unlike the binomial distribution, where $m = 2$). Additionally, they do not need to have equal probabilities: outcome i has probability p_i, so the sum of the p_is is 1. Then, X_i represents the number of trials with outcome i. Then,

$$P(X_1 = c_1, \ldots, X_m = c_m) = \binom{n}{c_1, \ldots, c_n} \prod_{i=1}^m p_i^{c_i} \qquad \text{where } \sum_{i=1}^m c_i = n.$$

Example 12.1. Suppose a six-sided die is rolled 7 times. The probability of one 1, one 2, no threes, two 4s, no 5s, and three 6s is

$$P(X_1=1,X_2=1,X_3=0,X_4=2,X_5=0,X_6=3)=\frac{7!}{1!1!0!2!0!3!}\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^3=\frac{420}{6^7}.$$

The multinomial distribution can be used in probabilistic text analysis: what is the probability that any given word appears in a text? In some sense, each word appears with some given probability, which makes this a giant multinomial distribution. This can be conditioned on the author, which makes sense: different people use different writing styles. But then, it can be flipped around: using Bayes' theorem, one can determine the probability of a writer for a given set of words. This was actually used on the Federalist Papers from early American history, which allowed for a probabilistic identification of which of Hamilton, Madison, and Jay wrote each paper. Similarly, one can use this to construct spam filters, as it allows one to guess whether the author of an email was a spammer.

Independence has already been discussed in the context of events, but it also applies to variables. The intuition is exactly the same: knowing the value of X indicates nothing about that of Y.

Definition. Two discrete random variables X and Y are indepedent if $p(x,y)=p_X(s)p_Y(y)$ for all x,y. Two variables that aren't independent are called dependent.

Example 12.2. Take a coin with a probability p of heads and flip it $m+n$ times. Then, let X be the number of heads in the first n flips and Y be the number of heads in the next m flips. It is not surprising that

$$P(X=x,Y=y)=\binom{n}{x}p^x(1-p)^{n-x}\binom{n}{y}p^y(1-p)^{m-y}=P(X=x)P(Y=y).$$

Thus, X and Y are independent.

More interestingly, let Z be the number of total heads in all $n+m$ flips. Then, X and Z aren't independent (e.g. $Z=0$), nor are Y and Z.

Example 12.3. Let $N\sim\text{Poi}(\lambda)$ be the number of requests to a web server in a day. If each request comes from a human with probability p and a bot with probability $1-p$, then let X be the number of humans per day. This means that $(X\mid N)\sim\text{Bin}(N,p)$. Similarly, if Y is the number of bot requests per day, then $(Y\mid N)\sim\text{Bin}(N,1-p)$. These can be made into a joint distribution:

$$\begin{aligned}P(x=i,Y=j)&=P(X=i,Y=i\mid X+Y=i+j)P(X+Y=i+j)+P(X=i,Y=i\mid X+Y\neq i+j)P(X+Y\neq i+j)\\&=P(X=i,Y=i\mid X+Y=i+j)P(X+Y=i+j).\end{aligned}$$

Then, $P(X=i,Y=j\mid X+Y=i+j)=\binom{i+j}{i}p^i(1-p)^j$ since it's multinomial, and $P(X+Y=i_j)=e^{-\lambda}\lambda^{i+j}/(i+j)!$, so the overall probability is

$$\begin{aligned}P(X=i,Y=j)&=\binom{i+j}{i}p^i(1-p)^je^{-\lambda}\frac{\lambda^{i+j}}{(1+j)!}\\&=e^{-\lambda}\frac{(\lambda p)^i}{i!}\frac{(\lambda(1-p))^j}{j!}\\&=e^{-\lambda p}\frac{(\lambda p)^i}{i!}e^{-\lambda(1-p)}\frac{(\lambda(1-p))^j}{j!}=P(X=i)P(Y=j),\end{aligned}$$

where $X\sim\text{Poi}(\lambda p)$ and $Y\sim\text{Poi}(\lambda(1-p))$.[18] Thus, X and Y are in fact independent, which makes sense.

Definition. Two continuous random variables X and Y are independent if $P(X\le a,Y\le b)=P(X\le a)P(Y\le b)$ for any a and b.

This is equivalent to any of the following:

- $F_{X,Y}(a,b)=F_X(a)F_Y(b)$ for all a,b.
- $f_{X,Y}(a,b)=f_X(a)f_Y(b)$ for all a,b.
- More generally, $f_{X,Y}(x,y)=h(x)g(y)$ with $x,y\in\mathbb{R}$, which has nice consequences for integration. Note that the constraints have to factor as well (see below).

Example 12.4.

- If $f_{X,Y}(x,y)=6e^{-3x}e^{-2y}$, then X and Y are pretty clearly independent: let $h(x)=3e^{-3x}$ and $g(y)=2e^{-2y}$.
- If $f_{X,Y}(x,y)=4xy$ for $0<x,y<1$, then X and Y are still independent, as $h(x)=2x$ and $g(y)=2y$.

[18] For some reason, this was a relevant xkcd today.

- With $f_{X,Y}$ as above, with the additional constraint that $0 < x + y < 1$. Here, X and Y aren't independent, both intuitively and since the constraints can't factor independently.

Example 12.5. Suppose two people set up a meeting at noon, and each arrives at a time uniformly distributed between noon and 12:30 p.m. Let X be the minutes past noon that the first person arrives, and Y be the minutes past noon person 2 arrives. Then, $X, Y \sim \text{Uni}(0, 30)$. What is the probability that the first person to arrive waits more than ten minutes for the other?

This problem is symmetric in X and Y, so $P(X + 10 < Y) + P(Y + 10 < X) = 2P(X + 10 < Y)$. Then,

$$\begin{aligned} 2P(X+10<Y) &= 2 \iint\limits_{x+10<y} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 2 \iint\limits_{x+10<y} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= 2\int_{10}^{30}\int_{0}^{y-10} \left(\frac{1}{30}\right)^2 \mathrm{d}x\,\mathrm{d}y = \frac{2}{30^2}\int_{10}^{30}\int_{0}^{y-10} \mathrm{d}x\,\mathrm{d}y = \frac{2}{30^2}\int_{10}^{30}(y-10)\,\mathrm{d}y = \frac{4}{9} \end{aligned}$$

The hardest part of this was probably finding the boundary conditions; the integration is straightforward, and some of the steps were skipped.

Example 12.6. Suppose a disk surface is a disc of radius R and a single point of imperfection is uniformly distributed on the disk. Then,

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi R^2}, & x^2 + y^2 \le R^2 \\ 0, & x^2 + y^2 > R^2 \end{cases}$$

Notice that X and Y, which are the random variables that are the coordinates of that point, are not independent: if $X = 1 = 0$. More formally,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{\pi R^2} \int\limits_{x^2+y^2\le R^2} \mathrm{d}y = \int\limits_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \mathrm{d}y = \frac{2\sqrt{R^2-x^2}}{\pi R^2}.$$

By symmetry, $f_Y(y) = 2\sqrt{R^2 - y^2}/\pi R^2$, where $-R \le y \le R$. Thus, $f_{X,Y}(x, y) \ne f_X(x) f_Y(y)$.

Often, the distance of this imperfection is important: more important data tends to be placed on the outside so that it can be accessed faster. If $D = \sqrt{X^2 + Y^2}$, then $P(D \le A) = \pi a^2/\pi R^2 = a^2/R^2$, since it can be modelled as a random dart landing in the circle of radius a.

Additionally, the expected value can be calculated, though it might be counterintuitive: using Lemma 11.1,

$$E[D] = \int_0^R P(D > a)\,\mathrm{d}a = \int_0^R \left(1 - \frac{a^2}{R^2}\right)\mathrm{d}a = \left[a - \frac{a^3}{3R^2}\right]_0^R = \frac{2R}{3}.$$

Independence of n variables for $n > 2$ is a straightforward generalization of the $n = 2$ case, since it requires the probabilities to factor through for any possible subset. This is identical to the definition for events. Indepenence is symmetric: if X and Y are independent, then Y and X are. This is obvious, but consider a sequence $X_1, X_2, \ldots$ of independent and identically distributed (IID)[19] Then, X_n is a record value if $X_n > X_i$ for all $i < n$ (i.e. $X_n = \max(X_1, \ldots, X_N)$). Let A_i be the event that X_i is a record value. Then, the probability that tomorrow is a record value seems to depend on whether today was, but flipping it around, does tomorrow being a record value affect whether today is?

More mathematically, $P(A_n) = 1/n$ and $P(A_{n+1}) = 1/(n + 1)$. Then, $P(A_n A_{n+1}) = (1/n)(1/(n + 1)) = P(A_n)P(A_{n+1})$.

Example 12.7. One useful application is choosing a random subset of size k from a set of n elements. The goal of this is to do this such that all $\binom{n}{k}$ possibilities are equally likely, given a `uniform()` function that simulates $\text{Uni}(0, 1)$. There's a brute-force way to do this by calculating all subsets and then picking a random one. This is exponential in time and space. Here's a better solution:

```
int indicator(double p) {
    if (random() < p) return 1; else return 0;
}
subset rSubset(k, set of size n) {
    subset_size = 0;
    I[1] = indicator((double)k/n);
    for(i = 1; i < n;i++) {
        subset_size += I[i];
```

[19] The second condition means that they all have the same distribution, not even depending on i.

```
            I[i+1] = indicator((k-subset_size)/(n-i));
        }
        return (subset containing element[i] iff I[i] == 1);
    }
```

13. Adding Random Variables: 4/29/13

Recall Example 12.7 from the previous lecture. The general idea of the algorithm is to create an array `I` with exactly k nonzero entries. Each element is added with a probability that depends on the number of elements left in the set. Also, note that this algorithm is linear, which is a nice speedup from the previous case.

Claim. Any given subset of k elements is equally likely to be returned from this algorithm.

Proof. Proceed by induction $k+n$. In the base case, where $k=1$ and $n=1$, then $S=\{a\}$ and `rSubset` returns $\{a\}$ with probability $p=1=1/\binom{1}{1}$.

In the general case, suppose the algorithm works for $k+x\le c$. Then:

- If $k+n\le c+1$, then $|S|=n$. Suppose that $\texttt{I}[1]=1$. Then, by the inductive hypothesis, `rSubset` returns a subset S' of size $k-1$ with probability $1/\binom{n-1}{k-1}$ (i.e. equally likely). Then, since $P(\texttt{I}[1]=1)=k/n$, then the total probability of any subset which contains the first element is $k/n\cdot 1/\binom{n-1}{k-1}=\binom{n}{k}$.
- If $k+n\le c+1$ and $\texttt{I}[1]=0$, then a set of size k is returned out of the remainin $n-1$ with probability $1/\binom{n-1}{k}$. Then, $P(\texttt{I}[1]=0)=(1-k/n)$, so the overall probability is $(1-k/n)(1/\binom{n-1}{k-1})=1/\binom{n}{k}$. ⊠

Suppose X and Y are independent random variables $X\sim\text{Bin}(n_1,p)$ and $Y\sim\text{Bin}(n_2,p)$. Then, $X+Y\sim\text{Bin}(n_1+n_2,p)$. Intuitively, do the trials for Y after those of X, and it's essentially the same variable, since both X and Y have the same chance of success per trial. This generalizes in the straightforward way: suppose $X_i\sim\text{Bin}(n_i,p)$ for $1\le i\le n$. Then,

$$\left(\sum_{i=1}^{n}X_i\right)\sim\text{Bin}\left(\sum_{i=1}^{n}n_i,p\right).$$

Now take $X\sim\text{Poi}(\lambda_1)$ and $Y\sim\text{Poi}(\lambda_2)$, where X and Y are independent. These can be summed as well: $X+Y\sim\text{Poi}(\lambda_1+\lambda_2)$. The key idea in the proof is that $P(X+Y=n)=P(X=k,Y=n-k)$ over all $0\le k\le n$. More generally, if $X_i\sim\text{Poi}(\lambda_i)$ are all independent, then

$$\left(\sum_{i=1}^{n}X_i\right)\sim\text{Poi}\left(\sum_{i=1}^{n}\lambda_i\right).$$

Then, some things can be said about density. Suppose that $g(X,Y)=X+Y$. Then, $E[g(X,Y)]=E[X+Y]=E[X]+E[Y]$. (The proof will be given in the next lecture.) This can be generalized to many variables $X_1,\dots,X_n$:

$$E\left[\sum_{i=1}^{n}X_i\right]=\sum_{i=1}^{n}E[X_i].$$

This holds *regardless of the dependencies between the* X_i. Thus, it is an incredibly useful result, and comes up all the time in probability (including the next few problem sets).

Then, consider two independent, continuous random variables X and Y. The CDF of $X+Y$ can be given by

$$F_{X+Y}(a)=P(X+Y\le a)=\iint_{x+y\le a}f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y=\int_{y=-\infty}^{\infty}\int_{x=-\infty}^{a-y}f_X(x)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y=\int_{-\infty}^{\infty}F_X(a-y)f_Y(y)\,\mathrm{d}y.$$

F_{X+Y} is called the convolution of F_X and F_Y as given above. This also works for PDFs:

$$f_{X+Y}(a)=\int_{-\infty}^{\infty}f_X(a-y)f_Y(y)\,\mathrm{d}y.$$

Convolutions exist in the discrete case, where integrals are replaced with sums and PDFs are replaced with probabilities.

Suppose $X\sim\text{Uni}(0,1)$ and $Y\sim\text{Uni}(0,1)$ are independent random variables (i.e. $f(a)=1$ for $0\le a\le 1$). Intuitively, the uniform distribution seems easy to handle, but behaves badly when something more complicated happens to it: here's what happens to the PDF of $X+Y$:

$$f_{X+Y}(a)=\int_0^1 f_X(a-y)f_Y(y)\,\mathrm{d}y=\int_0^1 f_X(a-y)\,\mathrm{d}y.$$

Now some casework is necessary: $0 \le a \le 2$, so consider $0 \le a \le 1$. Then, $0 \le y \le a$, so that $0 \le a - y \le 0$ and $f_X(a-y) = 1$, so $f_{X+Y}(a) = \int_0^a \mathrm{d}y = a$. However, if $1 \le a \le 2$, then $a - 1 \le y \le 1$, but $f_X(a-y) = 1$ still, so $f_{X+Y}(a) = \int_{a-1}^1 \mathrm{d}y = 2 - a$. Thus, the PDF is (as in Figure 5)

$$f_{X+Y}(a) = \begin{cases} a, & 0 \le a \le 1 \\ 2-a, & 1 < a \le 2 \\ 0, & \text{otherwise.} \end{cases}$$

Another fact (whose proof is a little complicated for now) is that if $X \sim \mathsf{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathsf{N}(\mu_2, \sigma_2^2)$ are independent, then

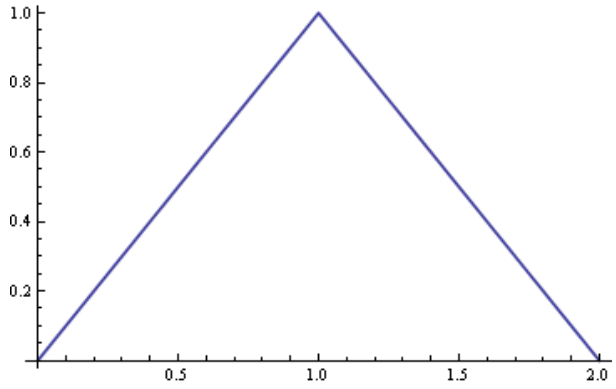


FIGURE 5. The sum of two uniform distributions on $(0, 1)$. Source

$X + Y \sim \mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, which is well-behaved even though it seems more complicated than the uniform distribution. This can be once again generalized: if $X_1, \dots, X_n$ are independent random variables, such that $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^n X_i = \mathsf{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2.\right)$$

Example 13.1. Suppose an RCC checks several computers for viruses. There are 50 Macs, each of which is independently infected with probabiliy $p = 0.1$ and 100 PCs, each independetly infected with probability $p = 0.4$. Let A be the number of infected Macs and B be the number of infected PCs, so $A \sim \mathsf{Bin}(50, 0, 1) \approx \mathsf{N}(5, 4.5)$, and $B \sim \mathsf{Bin}(100, 0.4) \approx \mathsf{N}(40, 24)$. Thus, $P(A + B \ge 40) \approx P(X + Y \ge 39.5)$. (Recall the continuity correction and inclusiveness discussed previously.) Since the normal distribution is additive, then $X + Y \sim \sim(45, 28.5)$, so $P(X + Y \ge 39.5) = 1 - \Phi(-1.03) \approx 0.8485$.

Conditional probabilities still work for distributions: the conditional PMF of X given Y (where $p_Y(y) > 0$) can be given by

$$P_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

This is not that different from before, but it is helpful to be clear about the difference between random variables and events. Since these are discrete random variables, sometimes breaking these into sums is helpful.

Example 13.2. Suppose a person buys two computers over time, and let X be the event that the first computer is a PC, and Y be the event that the second computer is a PC (as indicators, so $X, Y \in \{0, 1\}$). Then, various conditional probabilities can be found using the joint PMF table, which looks not too tricky. Interestingly, $P(X = 0 \mid Y = 1)$ can be calculated, which is sort of looking back in time. This can be used to make recommendations for future products, etc.

Example 13.3. Suppose a web server receives requests, where $X \sim \mathsf{Poi}(\lambda_1)$ is the number of requests by humans per day and $Y \sim \mathsf{Poi}(\lambda_2)$ is the number of requests from bots per day. Then, X and Y are independent, so they sum to

$\text{Poi}(\lambda_1 + \lambda_2)$. Then, given only the total number of requests one can get the probablity that k of them were by humans:

$$\begin{aligned} P(X = k \mid X + Y = n) &= \frac{P(X = k, Y = n - k)}{P(X + Y = n)} = \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} \\ &= \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{-(\lambda_1+\lambda_2)}(\lambda_1 + \lambda_2)^n} \\ &= \frac{n!}{k!(n-k)!} \cdot \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^{n-k}. \end{aligned}$$

This looks suspiciously like a binomial distribution, and $X \mid X + Y \sim \text{Bin}(X + Y, \lambda_1/(\lambda_1 + \lambda_2))$. Intuitively, for every request, a coin is flipped that determines the origin of each request.

If X and Y are continuous random variables, then conditional probability looks not much different: the conditional PDF is $f_{X|Y}(x \mid y) = f_{X,Y}(x, y)/f_Y(y)$. Intuitively, this is the limit of probabilities in a small area around x and y. Then, the CDF is

$$F_{X|Y}(a \mid y) = P(X \le a \mid Y = y) = \int_{-\infty}^{a} f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

Observe that it does make sense to condition on y even though $P(Y = a) = 0$, because this is actually a limit that does make sense. This would require chasing epsilons to show rigorously, however.

14. Continuous Conditional Distributions: 5/1/13

Though the definition of a conditional PDF was given in the last lecture, an example will make it make more sense.

Example 14.1. Suppose X and Y are continuous random variables with the PDF

$$f(x, y) = \begin{cases} \frac{12}{5}x(2 - x - y), & 0 < x, y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then, the conditional density is

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{f_{X,Y}(x, y)}{\int_0^1 f_{X,Y}(x, y)\,\mathrm{d}x} \\ &= \frac{(12/5)x(2 - x - y)}{\int_0^1 (12/5)(2 - x - y)\,\mathrm{d}x} = \frac{x(2 - x - y)}{[x^2 - x^3/3 - x^2y/2]_0^1} \\ &= \frac{6x(2 - x - y)}{4 - 3y}. \end{aligned}$$

Notice that this depends on both x and y, which makes sense.

If X and Y are continuous random variables, then $f_{X|Y}(x \mid y) = f_X(x)$, which is exactly the same as in the discrete case.

Suppose X is a continuous random variable, but N is a discrete random variable. Then, the conditional PDF of X given N is $f_{X|N}(x \mid n) = p_{N|X}(n \mid x)f_X(x)/p_N(n)$. Notice the use of PMFs for the discrete variable and PDFs for the continuous variable. What is slightly weirder is the conditional PMF of N given X, where the discrete distribution changes based on the continuous one: $p_{N|X}(n \mid x) = f_{X|N}(x \mid n)p_N(n)/f_X(x)$. Bayes' theorem happens to apply here as well.

One good example of this is a coin flip in which the probability of the coin coming up heads is unknown: it's a continuous random variable between 0 and 1. Then, the number of coins coming up heads is a discrete quantity that depends on a continuous one.

Definition. X is a beta random variable, denoted $X \sim \text{Beta}(a, b)$, if its PDF is given by

$$f(x) = \begin{cases} \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, & 0 < x < 1 \\ 0, & \text{otherwise,} \end{cases}$$

where $B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1}\,\mathrm{d}x$ is the normalizing constant.

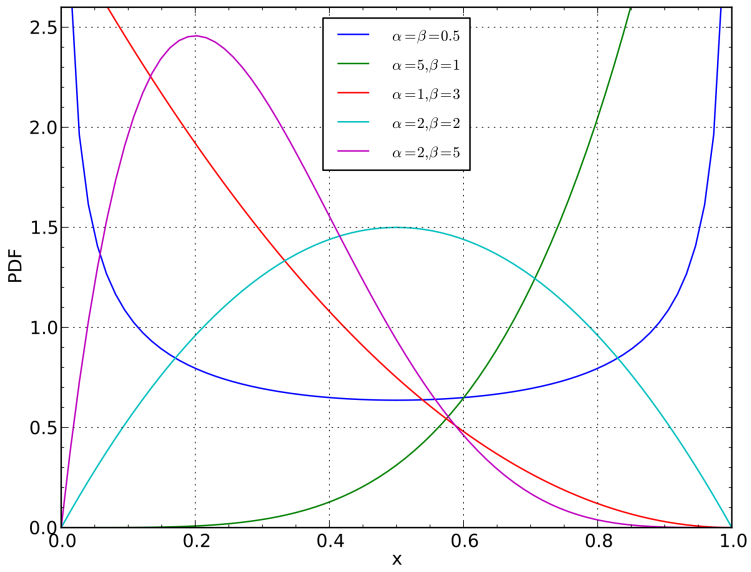


FIGURE 6. Several different examples of the beta distribution. Source

This distribution can do lots of unusual things, as in Figure 6, though when $a = b$ it is symmetric about $x = 0.5$. Then, $E[X] = a/(a+b)$ and $\operatorname{Var}(X) = ab/((a+b)^2(a+b+1))$.

This distribution is used as in the following example: suppose a coin is flipped $m+n$ times and comes up with n heads. Let X be the probability that the coin comes up heads (which is unknown), so that $X \sim \text{Uni}(0,1)$ and $f_X(x) = 1$. If X is known, then different coin flips are independent, but this isn't the case yet. Let N be the number of heads, so that $N \mid X \sim \text{Bin}(n+m, x)$. Then, the conditional density of X given that $N = n$ is

$$f_{X|N}(x \mid n) = \frac{P(N = n \mid X = x) f_X(x)}{P(N = n)} = \frac{\binom{n+m}{n} x^n (1-x)^m}{P(N = n)} = \frac{x^n (1-x)^m}{\int_0^1 x^n (1-x)^m \, \mathrm{d}x}.$$

The value on the bottom is obtained by a sneaky trick: there needs to be a normalizing constant, and $P(N = n)$ isn't known, so they have to match up. But this holds only when $0 < x < 1$, so $X \mid (N = n, n+m \text{ trials}) \sim \text{Beta}(n+1, m+1)$. In some sense, the more heads one obtains, the more the beta distribution tends towards 1, and the more tails, the more it skews to zero. As more and more trials are conducted, the beta distribution narrows in on the likely probability.

Another nice fact is that $\text{Beta}(1,1) = \text{Uni}(0,1)$ (verified by just calculating it out). It is also a conjugate distribution for itself (as well as the Bernoulli and binomial distributions) because the prior distribution (before a trial) of a trial given by the beta distribution makes calculating the posterior distribution (after the trial) easy (they're both beta distributions). In fact, one can set $X \sim \text{Beta}(a,b)$ as a guess as a prior to test whether a coin has a certain bias. This is called a subjective probability. After $n+m$ trials, the distribution can be updated: $X \mid (n \text{ heads}) \sim \text{Beta}(a+n, b+m)$. In some sense, the equivalent sample size, or the initial guess, is overwritten in the long run. It represents the number of trials (specifically, $a+b-2$ trials) imagined, of which $a-1$ ended up heads.

Returning to expectation, we know that if $a \le x \le b$, then $a \le E[X] \le b$, which is fairly easy to see formally. However, expectation can be generalized: if $g(X,Y)$ is a real-valued function, then in the discrete case

$$E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y),$$

since X and Y might not be independent (so their joint distribution is necessary). In the continuous case,

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) \, \mathrm{d}x \, \mathrm{d}y.$$

For example, if $g(X,Y) = X + Y$, then

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ \int_{-\infty}^{\infty}xf_X(x)\,\mathrm{d}x &+ \int_{-\infty}^{\infty}yf_Y(y)\,\mathrm{d}y = E[X]+E[Y], \end{aligned}$$

without any assumptions about X and Y (especially as to their independence).

Here's another reasonably unsurprising fact: if $P(a \le x) = 1$, then $a \le E[X]$. This is most often useful when x is nonnegative. However, the converse is untrue: if $E[X] \ge a$, then it isn't necessarily true that $X \ge a$ (e.g. if X is equally likely to take on -1 or 3, so that $E[X] = 1$). Similarly. if $X \ge Y$, then $X - Y \ge 0$, so $E[X-Y] = E[X] - E[Y] \ge 0$, so $E[Y] \ge E[Y]$. Again, the converse is untrue.

Definition. Let $X_1, \dots, X_n$ be independently and identically distributed random variables. If F is some distribution function and $E[X_i] = \mu$, then a sequence of X_i is called a sample from the distribution F. Then, the sample mean is $\overline{X} = \sum_{i=1}^{n} X_i/n$.

The idea is that $X_1, \dots, X_n$ are chosen from a distribution (students taking a test), and the mean is just that of this sample. The sample mean can vary depending on how the sample ends up. In some sense, $E[\overline{X}]$ represents the average score of a class on a test, the average height on a test, etc. $\overline{X}$ is a random variable. Then,

$$E[\overline{X}] = E\left[\sum_{i=1}^{n}\frac{X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu.$$

Thus, $\overline{X}$ is an "unbiased" estimate of μ, since $E[\overline{X}] = \mu$.

Let $E_1, \dots, E_n$ be events with indicator random variables X_i (i.e. $X_i = 1$ if E_i happens, and is zero otherwise). Then, $E[X_i] = P(E_i)$. Let $X = \sum X_i$ and $Y = 1$ if $X \ge 1$ and 0 otherwise. Then, $X \ge Y$, so $E[X] \ge E[Y]$, and $E[X] = E\left[\sum X_i\right] = \sum E[X_i] = \sum P(E_i)$, but $E[Y] = P\left(\bigcup_{i=1}^{n} E_i\right)$, so

$$\sum_{i=1}^{n} P(E_i) \ge P\left(\bigcup_{i=1}^{n} E_i\right).$$

This is known as Boole's inequality.[20]

If $Y \sim \mathrm{NegBin}(r, P)$, let X_i be the number of trials needed to get success after the $(i-1)^{\text{st}}$ success, so $X_i = \mathrm{Geo}(p)$ and $E[Y] = r/p$.

15. More Expectation: 5/3/13

The first revelation of today: apparently I look like Dr. Sahami. I don't see it, I guess.

Example 15.1. Consider a hash table with n buckets, and each string is equally likely to be hashed into each bucket.[21] Let X be the number of strings to hash until each bucket has at least one string. Finding $E[X]$ is a somewhat complicated problem, so break it into simpler ones. Let X_i be the number of trials to get success after the i^{th} success, where "success" means hashing a string into a previously empty bucket. After i buckets are nonempty, the probability of hashing a string into an empty bucket is $p = (n-i)/n$. Thus, $P(X_i = k) = ((n-i)/n)(i/k)^{k-1}$, so $X_i \sim \mathrm{Geo}((n-i)/n)$, so $E[X_i] = 1/p = n/(n-i)$. Then, $X = \sum_{i=0}^{n-1} X_i$, so

$$E[X] = \sum_{i=0}^{n-1} E[X_i] = \sum_{i=0}^{n-1}\frac{n}{n-i} = n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + 1\right) = O(n\log n).$$

Example 15.2. Recall Quicksort, a fast, recursive sorting algorithm. Choosing the index is based on a partition function:

```
int Partition(int[] arr, int n) {
    int lh = 1, rh = n-1;
    int pivot = arr[0];
    while(true) {
        while (lh < rh && arr[rh] >= pivot) rh--;
        while (lh < rh && arr[lh] < pivot) lh++;
```

[20] This is the same Boole who invented the notion of Boolean logic.

[21] This is an example of a coupon collecting model, though in CS applications it tends not to be called that.

```
            if (lh == rh) break;
            Swap(arr[lh],arr[rh]);
        }
        if (arr[lh] >= pivot) return 0;
        Swap(arr[0],arr[lh]);
        return lh;
    }
```

The complexity of this algorithm is determined by the number of comparisons made to the pivot. Though Quicksort is $O(n \log n)$, in the worst case it is $O(n^2)$, in which every time the pivot is selected, it is the maximal or minimal remaining element. Then, the probability that Quicksort has the worst-case behavior is the same as the probability of creating a degenerate BST (as seen on the first problem set), since there are exactly two bad pivots on each recursive call. Thus, the probability of the worst case is $2^{n-1}/n!$.

This isn't enough to give a nice description of the running time; the goal is to instead find the expected running time of the algorithm. Let X be the number of comparisons made when sorting n elements, so that $E[X]$ is the expected running time of the algorithm.

Let $X_1, \dots, X_n$ be the input to be sorted, and let $Y_1, \dots, Y_n$ be the output in sorted order. Let $I_{a,b} = 1$ if Y_a and Y_b are compared and 0 otherwise. Then, because of the order, each pair of values is only compared once: $X = \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} I_{a,b}$.

For any given Y_a and Y_b, if the pivot chosen is not between them, they aren't directly compared in that recursive call. Thus, the only cases we care about are those in which the pivot is in $\{Y_a, \dots, Y_b\}$. Then, in order to compare then, one of them must be selected as the pivot, so the probability that they are compared is $2/(b-a+1)$. This has the interesting consequence that randomly permuting an array can make Quicksort run faster.

Then, one can make some explicit calculations: first, an approximation is made:

$$\sum_{b=a+1}^{n} \frac{2}{b-a+1} \approx \int_{a+1}^{n} \frac{2\,\mathrm{d}b}{b-a+1} = 2\ln(n-a+1)\big|_{a+1}^{n} \approx 2\ln(n-a+1)$$

when n is large. Thus, let $y = n - a + 1$.

$$E[X] \approx \sum_{a=1}^{n-1} 2\ln(n-a+1) \approx 2\int_{1}^{n-1} \ln(n-a+1)\,\mathrm{d}a = -2\int_{n}^{2} \ln y\,\mathrm{d}y \approx 2n\ln(n) - 2n = O(n\log n).$$

Let I_i be indicator variables for events A_i. Then, we can speak of pairwise indicators: the variable $I_i I_j$ is the indicator for the event $A_i \cap A_j$. Thus, one has

$$E\left[\binom{X}{2}\right] = E\left[\sum_{i<j} I_j I_j\right] = \sum_{i<j} E[I_i I_j] = \sum_{i<j} P(A_i \cap A_j).$$

This can also be derived by expanding $\binom{X}{2} = x(x-1)/2$. Then, $E[X^2] - E[X] = 2\sum_{i<j} P(A_i \cap A_j)$, so $E[X^2] = 2\sum_{i<j} P(A_i \cap A_j) - E[X]$. Thus, one has the formula for variance, which can be used to obtain the variance of the binomial distribution another way.

Example 15.3. Consider a computer cluster with k servers. Requests independently go to each server i with probability p_i, and then, one can calculate how many machines are being useful: let A_i be the probability that server i receives no requests. Let X be the number of events A_i that occur, and $Y = k - X$ be the number of machines that do some work. Since the requests are indepdent, $P(A_i) = (1-p_i)^n$, so

$$E[Y] = k - E[X] = k - \sum_{i=1}^{k} P(A_i) = \sum_{i=1}^{k} (1-p_i)^n,$$

But knowing the variance is also helpful in terms of this real-world model. Since the events X_i are independent, then $P(A_i \cap A_j) = (1-p_i-p_j)^n$ for $i \neq j$, so

$$E[X(X-1)] = E[X^2] - E[X] = 2\sum_{i<j} P(A_i \cap A_j) = 2\sum_{i<j} (1-p_i-p_j)^n,$$

so the variance is

$$\begin{aligned}\operatorname{Var}(X) &= 2\sum_{i<j}(1-p_i-p_j)^n + E[X] - (E[X])^2 \\ &= 2\sum_{i<j}(1-p_i-p_j)^n + \sum_{i=1}^{k}(1-p_i)^n - \left(\sum_{i=1}^{k}(1-p_i)^n\right)^2 = \operatorname{Var}(Y).\end{aligned}$$

Places such as Amazon take their excess capacity and sell it to other people who might need servers, and knowing exactly how much they can sell is a valuable piece of information.

This example sounds exactly like coupon collecting or using hash tables, but is ostensibly more interesting.

One can take products of expectations: suppose X and Y are independent random variables and g and h are real-valued functions. Then, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$. If X and Y are dependent, this rule doesn't hold, but the sum rule still does. Here's why the product rule works in the independent case:

$$\begin{aligned}E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)\,\mathrm{d}x\int_{-\infty}^{\infty} h(y)f_Y(y)\,\mathrm{d}y \\ &= E[g(X)]E[h(Y)].\end{aligned}$$

Independence was necessary so that the joint PDF factored into the marginal density functions.

Definition. If X and Y are random variables, the covariance of X and Y is $\operatorname{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$.

Equivalently, $\operatorname{Cov}(X,Y) = E[XY - E[X]Y - XE[Y] + E[Y]E[X]] = E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y]$. This latter formula ($\operatorname{Cov}(X,Y) = E[XY] - E[X]E[Y]$) is particularly helpful, and indicates how X and Y vary together linearly. Clearly, if X and Y are independent, then $\operatorname{Cov}(X,Y) = 0$. However, the converse is not true, and this is a frequently misunderstood aspect of probability: if $X \in \{-1,0,1\}$ with equal likelihood and Y is an indicator for $Y = 0$, then $E[XY] = 0$, since $XY = 0$. Thus, $\operatorname{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0$, but X and Y aren't independent.

Example 15.4. Imagine rolling a six-sided die and let X be an indicator for returning a 1, 2, 3, or 4, and let Y be an indicator for 3, 4, 5, or 6. Then, $E[X] = E[Y] = 2/3$ and $E[XY] = 1/3$. Thus, the covariance is $\operatorname{Cov}(X,Y) = -1/9$, which is negative. This is surprising: the variance can't be negative, but this is, indicating that X and Y vary in opposite directions.

16. Covariance: 5/6/13

"It's not 'second hat,' it's 'old hat.' It's a mixed metaphor, like 'that's the way the cookie bounces.' "

Another example of covariance is a table of weights and heights. In a sample mean of several people's weights and heights, there is a significant covariance between them.

Here are some properties of covariance:

- $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$. This is not hard to see from the definition.
- $\operatorname{Cov}(X,X) = E[X^2] - E[X]E[X] = \operatorname{Var}(X)$.
- $\operatorname{Cov}(aX+bY) = a\operatorname{Cov}(X,Y)$. Notice that this is different than the variance, where $\operatorname{Var}(aX) = a^2\operatorname{Var}(X)$.
- If $X_1,\dots,X_n$ and $Y_1,\dots,Y_m$ are all random variables, then

$$\operatorname{Cov}\left(\sum_{i=1}^{n}X_i, \sum_{i=1}^{m}Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\operatorname{Cov}(X_i,Y_j).$$

This allows us to understand the variance of a sum of variables: it isn't just the sum of the variances.

Claim. If $X_1,\dots,X_n$ are random variables, then

$$\operatorname{Var}\left(\sum_{i=1}^{n}X_i\right) = \sum_{i=1}^{n}\operatorname{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\operatorname{Cov}(X_i,X_j).$$

Proof. Because $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$, then

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \operatorname{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \operatorname{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

The last step was done because $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$, so the sum can be simplified. ⊠

Note that this formula means that if X_i and X_j are independent when $i \neq j$ (since X_i can't be independent with itself), then $\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i)$. This is an important difference between the expected value and the variance.

Example 16.1. Let $Y \sim \text{Bin}(n, p)$ and let X_i be an indicator for the i^{th} trial being successful. Specifically, $X_i \sim \text{Ber}(p)$ and $E[X_i] = p$. Notice that $E[X^2] = E[X]$, since X is only 0 or 1, each of which squares to itself. Then,

$$\operatorname{Var}(Y) = \sum_{I=1}^{n} \operatorname{Var}(X_i) = \sum_{i=1}^{n} (E[X_i^2] - E[X_i]^2) = \sum_{i=1}^{n} (E[X_i] - E[X_i]^2) = \sum_{i=1}^{n} (p - p^2) = np(1-p).$$

This is a fact that you should have already seen before, but this is an interesting derivation.

Recall the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ for independently and identically distributed random variables $X_1, \ldots, X_n$. If $E[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2$, then $E[\overline{X}] = \mu$, but the variance is slightly more interesting:

$$\begin{aligned}\operatorname{Var}(\overline{X}) &= \operatorname{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \left(\frac{1}{n}\right)^2 \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \operatorname{Var}(X_i) \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

The variance gets smaller as there are more variables in the mean. This makes sense: as more things (e.g. people's heights) are added to the sample, then they on average look like the mean, so the variance decreases.

One can also define the sample deviation $\overline{X} - X_i$ for each i, and then obtain the sample variance, which is the guess at what the standard deviation should be:

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}.$$

Then, $E[S^2] = \sigma^2$. This means that S^2 is an "unbiased estimate" of σ^2, which is a bit of jargon. The $n-1$ in the denominator comes out after an intimidating computation to show that $E[S^2] = \sigma^2$.

Definition. If X and Y are two arbitrary random variables, the correlation of X and Y is

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}.$$

In some sense, this takes the covariance, which already illustrates something similar, and divides out to account for possibly different units. The correlation is always within the range $[-1, 1]$. This measures how much X and Y vary in a linear relationship: if $\rho(X, Y) = 1$, then $Y = aX + b$, where $a = \sigma_y/\sigma_X$ (the standard deviations of Y and X, respectively). If $\rho(X, Y) = -1$, then $Y = aX + b$, but $a = -\sigma_y/\sigma_X$. If $\rho(X, Y) = 0$, then there is no linear relationship between the two. Note that there may still be a relationship, just of a higher order. However, the linear component is typically the one that people care the most about.

If $\rho(X, Y) = 0$, then X and Y are uncorrelated; otherwise, they are correlated. Notice that independence implies two variables are uncorrelated (since the covariance is zero), but the converse is false!

Suppose I_A and I_B are indicator variables for events A and B. Then, $E[I_A] = P(A)$ and $E[I_B] = P(B)$, so $E[I_A I_B] = P(A \cap B)$. Thus,

$$\begin{aligned}\operatorname{Cov}(I_A, I_B) &= E[I_A I_B] - E[I_A]E[I_B] = P(A \cap B) - P(A)P(B) \\ &= P(A \mid B)P(B) - P(A)P(B) \\ &= P(B)(P(A \mid B) - P(A)).\end{aligned}$$

Thus, if $P(A \mid B) > P(A)$, then the correlation is positive; when A happens, then B is more likely. Similarly, if $P(A \mid B) = P(A)$, then $\rho(I_A, I_B) = 0$, and if $P(A \mid B) < P(B)$, then observing A makes B less likely, so they are negatively correlated.

Example 16.2. Suppose n independent trials of some experiment are performed, and each trial results in one of m outcomes, each with probabilities $p_1, \dots, p_m$, where the p_i sum to 1. Let X_i be the number of trials with outcome i. For example, one could count the number of times a given number comes up on a die when rolled repeatedly. Intuitively, two numbers should be negatively correlated, since more of one number implies less room, so to speak, for another.

Let $I_i(k)$ be an indicator variable that trial k has outcome i. Then, $E[I_i(k)] = p_i$, $X_i = \sum_{k=1}^{n} I_i(k)$. If $a \neq b$, then trials a and b are independent, so $\operatorname{Cov}(I_i(b), I_j(a)) = 0$, and when $a = b$, then $E[I_i(a)I_j(a)] = 0$ (you can't roll both a 1 and a 2 in the same trial), so $\operatorname{Cov}(I_i(a), I_j(a)) = -E[I_i(a)]E[I_j(a)]$, since the first term drops out. Thus,

$$\operatorname{Cov}(X_i, X_j) = \sum_{a=1}^{n}\sum_{a=1}^{n} \operatorname{Cov}(I_i(b), I_j(a)) = \sum_{a=1}^{n} \operatorname{Cov}(I_i(a), I_j(a)) = \sum_{a=1}^{n} -E[I_i(a)]E[I_j(a)] = \sum_{i=1}^{n} -p_i p_j = -np_i p_j,$$

so they are in fact negatively correlated. This is important because multinomial distributions happen in many cases in applications. For example, if m is large, so that p_i is small, and the outcomes are equally likely (such that $p_i = 1/m$), then $\operatorname{Cov}(X_i, X_j) = -n/m^2$, so the covariance decreases quadratically as m increases. This is why the Poisson paradigm works so well: this quantifies how accurate the approximation can be.

Suppose that X and Y are jointly discrete random variables, so their conditional PMF is $p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = p_{X,Y}(X, y)/p_Y(y)$. Then, the conditional expectation of X given that $Y = y$ is

$$E[X \mid Y = y] = \sum_x xP(X = x \mid Y = y) = \sum_x p_{X|Y}(x \mid y),$$

and in the continuous case we have the analogous result:

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

Example 16.3. Roll two six-sided dice D_1 and D_2. Then $E[X \mid Y = 6] = \sum_x xP(X = x \mid Y = 6) = (1/6)(7 + \cdots + 12) = 9.5$. This makes sense, as it is also $E[X] + Y$.

Example 16.4. If X and Y are independent random variables, such that $X, Y \sim \mathrm{Bin}(n, p)$, to compute $E[X \mid Y = m]$ it will be necessary to compute $P(X = k \mid X + Y = m)$:

$$\begin{aligned}P(X = k \mid X + Y = m) &= \frac{P(X = k, X + Y = m)}{P(X + Y = m)} = \frac{P(X = k, Y = m - k)}{P(X + Y = m)} = \frac{P(X = k)P(Y = m - k)}{P(X + Y = m)} \\ &= \frac{\binom{n}{k}p^k(1-p)^{n-k} \cdot \binom{n}{m-k}p^{m-k}(1-p)^{n-(m-k)}}{\binom{2n}{m}p^m(1-p)^{2n-m}} = \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}.\end{aligned}$$

This is a hypergeometric distribution: $(X \mid X + Y = m) \sim \mathrm{HypG}(m, 2n, n)$. This actually has an intuitive explanation: in a total of m draws, there are $2n$ total balls in an urn, of which n are white. This is exactly what the definition specifies. Thus, the expectation is already known: $E[X \mid X + Y = m] = nm/(2n) = m/2$. Of course, by symmetry, this makes sense, and you could even write it down ahead of time with some foresight.

The sum of expectations still holds, no matter what you throw at it: if $X_1, \dots, X_n$ and Y are random variables, then

$$E\left[\sum_{i=1}^{n} X_i \mid Y = y\right] = \sum_{i=1}^{n} E[X_i \mid Y = y].$$

Since $E[E \mid Y = y]$ sums over all values of X, then it can be thought of as a function of Y but not X, so $g(Y) = E[X \mid Y]$ is a random variable. For any given $Y = y$, $g(Y) = E[X \mid Y = y]$. Thus, we can take $E[g(Y)] = E[E[X \mid Y]]$. Intuitively, this is the expectation of an expectation, which simplifies easily, but these are expectations of different variables, so

they do *not* cancel out. The notation might be confusing, so add subscripts corresponding to the expected variables if necessary. However, in this case the expectation can be taken one layer at a time:

$$\begin{aligned}
E[E[X \mid Y]] &= \sum_y E[X \mid Y = y]P(Y = y) \\
&= \sum_y \left(\sum_x xP(X = x \mid Y = y) \right) P(Y = y) \\
&= \sum_{x,y} xP(X = x, Y = y) = \sum_x x \sum_y P(X = x, Y = y) \\
&= \sum_x xP(X = x) = E[X].
\end{aligned}$$

By replacing the sums with integrals, the same result can be found for the continuous case.

Example 16.5. Consider the following code, which recurses probabilistically:

```
int Recurse() {
    int x = randomInt(1,3); //equally likely values
    if (x == 1) return 3;
    else if (x == 2) return (5 + Recurse());
    else return (7 + Recurse());
}
```

Let Y be the value returned by `Recurse()`. What is $E[Y]$? This may seem like a weird example, but it could indicate the running time of a snippet of code given some different recursive calls.

Notice that $E[Y \mid X = 1] = 3$, $E[Y \mid X = 2] = 5 + E[Y]$, and $E[Y \mid X = 3] = 7 + E[Y]$, where X is the value of `x` in the program. Thus, $E[Y] = 3(1/3) + (5 + E[Y])/3 + (7 + E[Y])/3$, which can be solved for $E[Y] = 15$.

17. Predictions: 5/8/13

"Once the blue lightning comes out of [my daughter's] fingers, it's all over."

This example will exhibit a random number of random variables.

Example 17.1. Consider a website `www.PimentoLoaf.com`. Let $X_i \sim \mathrm{N}(50, 25)$ be the number of people who visit per day, and let $Y_i \sim \mathrm{Poi}(8)$ be the number of minutes spent by visitor i. X and the Y_i are all independent, so what is the expected number of minutes W that people spend on the site per day?

$$E[W] = E\left[\sum_{i=1}^{X} Y_i\right] = E\left[E\left[\sum_{i=1}^{X} Y_i \mid X\right]\right].$$

If it is known that $X = n$, then this becomes

$$E[W] = E\left[\sum_{i=1}^{n} Y_i \mid X = n\right] = \sum_{i=1}^{n} E[Y_i \mid X = n] = \sum_{i=1}^{n} E[Y_i] = nE[Y_i]$$

because the Y_i are identically distributed and independent of X, and thus the overall expectation becomes $XE[Y_i]$.

An important application of probability is to make predictions about some random variable Y given some other random variable X. This has lots of applications, such as high-frequency trading or prediction of genetic-based diseases. Let $g(X)$ be the function used to predict Y: the predicted value is $\hat{Y} = g(X)$. $g(X)$ is chosen so as to minimize $E[(Y - g(X))^2]$, which ends up being $g(X) = E[Y \mid X]$, which isn't actually a surprise. Intuitively, $E[(Y - c)^2]$ is minimized when $c = E[Y]$, and if X is observed, then let $c = E[Y \mid X]$. Under some relatively mild assumptions, it's not actually possible to do better. This is the beginning of machine learning.

Example 17.2. Suppose a father has height $X = 71$ inches, and historically, a son's height is $Y \sim \mathrm{N}(X + 1, 4)$ (i.e. $Y = (X+1)+C$, where $C \sim \mathrm{N}(4)$). Then, the predicted value for the height of the son is $E[Y \mid X = 71] = E[X+1+C \mid X = 71] = E[72 + C] + E[72] + E[C] = 72$.

Thus, one can compute probabilities by conditioning: if X is an indicator variable for the event A, then $E[X] = P(A)$ so $E[X \mid Y = y] = P(A \mid Y = y)$ for any Y. Thus, $E[X] = E_Y[E_X[X \mid Y]] = E_Y[P(A \mid Y)]$. In the discrete case, this

becomes $E[X] = \sum_y P(A \mid Y = y)P(Y = y) = P(A)$ by marginalizing.[22] More generally, if F_i is the indicator variable for $Y = y_i$, then

$$P(A) = \sum_{i=1}^{n} P(A \mid F_i)P(F_i),$$

which is called the law of total probability.

Example 17.3. Suppose n candidates for a job are interviewed for one position, and there is a factor of α (for some α) difference in productivity between the best and average candidate.[23] Imagine in this scenario a hiring decision has to be made immediately after the interview, which gives a ranking of the candidates that have already been interviewed.

One strategy is to first interview k candidates, and then hire the next candidate better than all of the first k candidates. Then, it's possible to calculate the probability that the best person is hired. Let X be the position of the best candidate in the line of interviews. Then, $P_k(\text{Best} \mid X = i) = 0$ if $i \leq k$. This is a good argument for making k small. In general, the best person of the first k sets the bar for hiring, so the best candidate in position i will be selected if the best of the first $i - 1$ is in the first k interviewed. Thus, if $i > k$, then $P_k(\text{Best} \mid X = i) = k/(i - 1)$. Then,

$$P_k(\text{Best}) = \frac{1}{n}\sum_{i=1}^{n} P_k(\text{Best} \mid X = i) = \frac{1}{n}\sum_{i=k+1}^{n} \frac{k}{i=1} \approx \frac{k}{n}\int_{k+1}^{n} \frac{\mathrm{d}i}{i-1} = \frac{k}{n}\ln\left(\frac{n}{k}\right).$$

Then, to optimize the value of k, differentiate this and set it equal to zero. This yields $k = n/e$, for which the probability of hiring the best person is about $1/e \approx 37\%$. This is interesting because it doesn't depend on the number of people interviewed.

Nonetheless, this isn't optimal, which is why most companies don't use this strategy.

Definition. The moment-generating function (MGF) of a random variable X is $M(t) = E\left[e^{tX}\right]$ where $t \in \mathbb{R}$.

When X is discrete, this is just $M(t) = \sum_x e^{tx}p(x)$, and when X is continuous, this is instead $M(t) = \int_{-\infty}^{\infty} e^{tx}f_X(x)\,\mathrm{d}x$. Interesting things happen when you calculate $M'(0)$: derivatives commute with taking the expected value, so $M'(X) = E[Xe^{tX}]$, so $M'(0) = E[X]$. Taking the derivative again yields $M''(t) = E[X^2e^{tX}]$, so $M''(0) = E[X^2]$, or the second moment. This generalizes: $M^{(n)}(0) = E[X^n]$.

Example 17.4. If $X \sim \text{Ber}(p)$, then $M(t) = E[e^{tX}] = e^0(1 - p) + e^tp = e^tp - p$, so $M'(0) = p$ and $M''(t) = e^tp$ again, so $M''(0) = p$ again. This is stuff we already knew, but it's nice to see it derived in a new way.

Example 17.5. Now let $X \sim \text{Bin}(n, p)$. Then,

$$M(t) = E\left[e^{tX}\right] = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=0}^{n}\binom{n}{k}(pe^t)^k(1-p)^{n-k} = (pe^t + 1 - p)^n$$

using the Binomial Theorem. Thus, $M'(t) = n(pe^t + 1 - p)^{n-1}pe^t$, so $M'(0) = E[X] = np$.[24] Then, in a similar process, one can show that $M''(0) = n(n - 1)p^2 + np$, so $\text{Var}(X) = np(1 - p)$.

If X and Y are independent random variables, then $M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$, and conversely, if the MGF factors, then these two variables are independent. Note that because of the exponential, the sum becomes a product. Additionally, $M_X(t) = M_Y(t)$ iff $X \sim Y$.

One can generalize to joint MGFs: if $X_1, \ldots, X_n$ are random variables, then their joint MGF is $M(t_1, \ldots, t_n) = E[e^{t_1X_1+\cdots+t_nX_n}]$. Then, $M_{X_i}(T) = M(0, \ldots, 0, t, 0, \ldots, 0)$, where the t is in the i^{th} position. Then, $X_1, \ldots, X_n$ are independent iff $M(t_1, \ldots, t_n) = \prod_{i=1}^{n} M_{X_i}(t)$, and the proof is the same as in the $n = 2$ case.

If $X \sim \text{Poi}(\lambda)$, then

$$M(t) = E[e^{tX}] = \sum_{n=0}^{\infty} e^{tn}\frac{e^{-\lambda}\lambda^n}{n!} = e^{\lambda(e^t-1)},$$

so after a bunch of somewhat ugly math, $M'(0) = \lambda$ and so on, as we already saw, but in a new way. This is significant because if $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$ are independent, this offers a simple proof that $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$.

[22] This also holds in the continuous case.

[23] For software engineers, Steve Jobs claimed $\alpha \approx 25$, but Mark Zuckerberg claims that $\alpha \approx 100$.

[24] This is… the seventh such derivation for this?

18. Moment-Generating Functions: 5/10/13

"In personality, consider an equilateral triangle with Sheen, his character, and myself at the three vertices."
- Herman Chernoff

Recall the definition of moment-generating functions from the previous lecture, which allowed one to obtain the n^{th} moment from the n^{th} derivative of the function. The MGF of a normal distribution $X \sim \mathsf{N}(\mu_1, \sigma_1^2)$ is

$$M_X(t) = E\left[e^{tX}\right] = e^{\left(\frac{\sigma_1^2 t^2}{2} + \mu_1 t\right)}.$$

Then, the first derivative is $M_X'(t) = (\mu_1 + t\sigma_1^2)M(t)$, so $M_X'(0) = \mu_1$, and after a bit more math, $M_X''(0) = \mu_1^2 + \sigma_1^2$. This makes sense. As for why it's actually important, suppose $Y \sim \mathsf{N}(\mu_2, \sigma_2^2)$ is independent of X.[25] Then, Y has the same MGF with indices changed, but one can calculate the convolution:

$$M_X(t)M_Y(t) = e^{\left(\frac{\sigma_1^2 t^2}{2} + \mu_1 t\right)} e^{\frac{\sigma_2^2 t^2}{2} + \mu_2 t} = e^{\left(\frac{(\sigma_1^2 + \sigma_2^2)t^2}{2} + (\mu_1 + \mu_2)t\right)} = M_{X+Y}(t),$$

so the normal distribution is additive: $X + Y \sim \mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. This has other meanings beyond the scope of this class, such as uses in Fourier analysis.

Here's another question: if $V = X + Y$ and $W = X - Y$, where $X, Y \sim \mathsf{N}(\mu, \sigma^2)$ are independent, are V and W independent? Take the joint MGF of V and W:

$$\begin{aligned} M(t_1, t_2) &= E\left[e^{t_1 V} e^{t_2 W}\right] = E\left[e^{t_1(X+Y)} e^{t_1(X-Y)}\right] = E\left[e^{(t_1+t_2)X} e^{(t_1-t_2)Y}\right] \\ &= E\left[e^{(t_1+t_2)X}\right] E\left[e^{(t_1-t_2)Y}\right] \text{ because } X \text{ and } Y \text{ are independent.} \\ &= e^{\mu(t_1+t_2) + \frac{\sigma^2(t_1+t_2)^2}{2}} e^{\mu(t_1-t_2) + \frac{\sigma^2(t_1-t_2)^2}{2}} \\ &= e^{2\mu t_1 + 2\sigma^2 t_1^2/2} e^{2\sigma^2 t_2^2/2} = E\left[e^{t_1 V}\right] E\left[e^{t_2 B}\right], \end{aligned}$$

where the last step can be seen by writing down the distributions for V and W. Notice that this result doesn't hold when X and Y aren't identically distributed, since the like terms can't be gathered.

In many cases, the true form of a probability distribution encountered in the real world isn't obvious. For example, this class' midterm didn't appear to be normally distributed. However, certain information (e.g. mean, sample variance) can be computed anyways, and certain properties can be seen by looking at the application of the problem (here, midterm scores are known to be nonnegative).

Proposition 18.1 (Markov's[26] Inequality). *Suppose X is a nonnegative random variable. Then, for any $a > 0$, $P(X \geq a) \leq E[X]/a$.*

Proof. Let I be an indicator for the event that $X \geq a$. Since $X \geq 0$, then $I \leq X/a$: if $X < a$, then $I = 0 \leq X/a$, and if $X \geq a$, then $X/a \geq 1 = I$. Then, take expectations: $E[I] = P(X \geq a) \leq E[X/a] = E[X]/a$, since constants factor out of expectation. ⊠

Applying this to the midterm, let X be a score on the midterm. Then, the sample mean is $\overline{X} = 95.4 \approx E[X]$, so $P(X \geq 110) \leq E[X]/110 = 95.4/110 \approx 0.8673$. This says that at most 86.73% of the class got greater than a 110. Since 27.83% of the class did this, it's clear that Markov's inequality is a very loose bound, but that makes sense because the only thing it knows is the mean.

Proposition 18.2 (Chebyshev's[27] Inequality). *Suppose X is a random variable such that $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Then, $P(|X - \mu| \geq k) \leq \sigma^2/k^2$ for all $k > 0$.*

Proof. Since $(X - \mu)^2$ is a nonnegative random variable, then apply Markov's inequality with $a = k^2$, so $P((X - \mu^2) \geq k^2) \leq E[(X - \mu)^2]/k^2 = \sigma^2/k^2$. However, $(X - \mu^2) \geq k^2$ iff $|X - \mu| \geq k$, so the inequality follows. ⊠

This is what is known as a concentration inequality, since it puts bounds on how spread the data can be from the mean. Using this for the midterm, the sample mean is $\overline{X} = 95.4 \approx E[X]$, and the sample variance is $S^2 = 400.40 \approx \sigma^2$, so $P(|X - 95, 4| \geq 22) \leq \sigma^2/22^2 \approx 0.8273$, so at most 82.73% of the class was more than 22 points away from the mean. This can also be used to calculate how much of the class is within a distance from the mean: at least 17.27% of the class

[25] μ_2 not to be confused with μ_2.

[26] Andrey Andreyevich Markov was a Russian mathematician who was responsible for this inequality as well as Markov chains, which are the basis of Google's PageRank algorithm.

[27] Pafnuty Lvovich Chebyshev was also a 19th-Century Russian mathematician who didn't actually formulate the inequality named after him. Interestingly enough, he was Markov's doctoral advisor and may have actually formulated Markov's inequality before Markov.

is within 22 points of the mean. Once again, this is very loose: about 83% of the class was within 22 points of the mean. Chebyshev's inequality is most useful as a theoretical tool.

There's another formulation called the one-sided Chebyshev inequality. In its simplest form, suppose that X is a random variable with $E[X] = 0$ and $\mathrm{Var}(X) = \sigma^2$. Then, $P(X \geq a) \leq \sigma^2/(\sigma^2 + a^2)$ for any $a > 0$. This generalizes: if $E[Y] = \mu$ and $\mathrm{Var}(Y) = \sigma^2$, then $P(Y \geq \mu + a) \leq \sigma^2/(\sigma^2 + a^2)$ and $P(Y \leq \mu - a) \leq \sigma^2/(\sigma^2 + a^2)$ for any $a > 0$. This follows directly from the first version by setting $X = Y - \mu$.

Applying this, at most 80.02% of the class could have gotten more than a 105.4, which is a pretty loose bound again, since 38.70% of the class did so. Nonetheless, it's interesting that something can be said with so little, and this is a better bound than Markov's inequality, which reports 0.9051 as an even looser bound.

Proposition 18.3 (Chernoff[28] bounds). *Let $M(t)$ be an MGF for a random variable X. Then, $P(X \geq a) \leq e^{-ta}M(t)$ for all $t > 0$ and $P(X \leq a) \leq e^{-ta}M(t)$ for all $t < 0$. This holds for all nonzero t, so one can pick the value that minimizes the bound.*

Proof. Using Markov's inequality, $P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq E[e^{tX}]/e^{ta} = e^{-ta}E[e^{tX}] = e^{-ta}M(t)$, and the other bound is about the same. ⊠

If $Z \sim \mathrm{N}(0, 1)$ (i.e. the standard normal), then its MGF is $M_Z(t) = e^{t^2/2}$, so the Chernoff bounds are $P(Z \geq a) \leq e^{ta}e^{t^2/2} = e^{t^2/2-ta}$ for all $t > 0$. Using calculus, the minimum value of this function is $t = a$, so $P(Z \geq a) \leq e^{-a^2/2}$ for all $a > 0$, and $P(Z \leq a) \leq e^{-a^2/2}$ for all $a < 0$. This is nice because it provides an approximation despite the fact that the PDF of the normal distribution isn't integrable.

In some sense, this reasonably tight bound exists because there's a one-to-one correspondence between the MGF and the distribution. However, one doesn't always know the MGF, so this is once again a theoretical tool in a lot of cases.

If $X \sim \mathrm{Poi}(\lambda)$, the MGF is $M_X(t) = e^{\lambda(e^t-1)}$, so the Chernoff bound is $P(X \geq i) \leq e^{\lambda(e^t-1)}e^{-it} = e^{\lambda(e^t-1)-it}$ for any $t > 0$. Thus, the goal is to minimize $\lambda(e^\lambda - 1) - it$ with respect to t, so that $e^t = i/\lambda$. Then, if $t > 0$, so that $i/\lambda > 1$, then $P(X \geq i) \leq e^t(i/\lambda - 1)(i/\lambda)^i = (e^\lambda/i)^i e^{-\lambda}$. Notice that this matches fairly closely to the functional form of the PMF.

Definition. A function f is convex if $f''(x) > 0$ for all $x \in \mathbb{R}$. f is concave if $-f$ is convex.

Intuitively, a convex function is bowl-shaped: functions such as $y = x^2$ and $y = e^x$ are convex.

Proposition 18.4 (Jensen's[29] inequality). *If f is a convex function, then $E[f(X)] \geq f(E[X])$.*

The proof of this is a fairly unpleasant Taylor expansion around μ. Intuitively, if f is convex, then it looks sort of like a bowl, and the line between two points on the graph of f is always above the graph itself. The expected value of $f(x)$ lies on this line, and $f(E[X])$ lies on the graph, so the former is greater than the latter. Similarly, if $f(E[X]) = E[f(X)]$, then they must both lie on the same line, so f is a line, since $f''(x) = 0$.

This has applications in utility theory, a branch of finance or economics. The utility $U(x)$ is the value that one derives from some object x. For example, one could play a game in which one wins \$20000 with a 50% chance, or receive \$10000 for not playing. Most people would decline to play in such a situation, even though the expected values are the same, so the first and second \$10000 have a different utility. Sometimes, the utility is nonmonetary (e.g. quality of life).

The notion of a utility curve is also helpful, in which x (here representing quantities of money) is graphed against $U(x)$. A concave utility curve indicates a risk-averse strategy: the first \$10000 has greater utility than subsequent ones, and a flat curve indicates there is no change (risk-neutral). However, a concave curve has a risk-preferring strategy. Then, Jensen's inequality says that if one can invest in something that always returns μ (some sort of bonds) or something which returns a value X such that $E[X] = \mu$ and $\mathrm{Var}(X) \neq 0$ (represented by stocks), then let R be the return value. Then, the goal is to maximize $U(R)$. The inequality says that a convex utility curve should invest more in stocks, and a concave utility curve should prefer bonds.

19. Laws of Large Numbers and the Central Limit Theorem: 5/13/13

Proposition 19.1 (Weak Law of Large Numbers). *Suppose $X_1, X_2, \ldots$ are independently and identically distributed random variables, with a distribution F, such that $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $\overline{X} = \sum_{i=1}^n X_i/n$. Then, for any $\varepsilon > 0$, $\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \varepsilon) = 0$.*

Proof. Since $E[\overline{X}] = \mu$ and $\mathrm{Var}(X) = \sigma^2/n$, then by Chebyshev's inequality, $P(|\overline{X} - \mu| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2)$, so as $n \to \infty$, this probability goes to zero. ⊠

[28] Herman Chernoff isn't Russian, and is still a Professor Emeritus at Harvard and MIT. He did discover Chernoff's Bound. Since he's still around, someone in 109 last quarter actually emailed him and asked him if he was a Charlie Sheen fan. He was.

[29] Johan Ludwig William Valdemar Jensen was a Dutch mathematician who was responsible for the inequality named after him.

Proposition 19.2 (Strong Law of Large Numbers). *With $X_1, X_2, \ldots$ and $\overline{X}$ as before,*

$$P\left(\lim_{n\to\infty}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\mu\right)=1.^{30}$$

Notice that the Strong Law implies the Weak Law, but not vice versa: the Strong Law says that for any $\varepsilon>0$, there are only a finite number of values n such that $|\overline{X}-\mu|\geq\varepsilon$ holds. The proof of the Strong Law isn't all that easy, and has been omitted.

Consider some set of repeated trials of an experiment, and let E be some outcome. Then, let X_i be an indicator for trial i returning E. Then, the Strong Law implies that $\lim_{n\to\infty}E[X]=P(E)$, which justifies more rigorously the fact shown in the first week of class, that $P(E)=\lim_{n\to\infty}n(E)/n=P(E)$. These laws are mostly philosophical, but they provide a mathematical definition of probability.

One misuse of the laws of large numbers is to justify the gambler's fallacy, in which after multiple losses of some sort, one expects a win. This is incorrect, because the trials are independent (and the distribution is memoryless anyways), and the averaging out happens as $n\to\infty$, which is not nearly small enough to fit in a lifetime.

The following result is one of the most useful in probability theory:

Theorem 19.3 (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a set of independently and identically distributed random variables with $E[X_i]=\mu$ and $\mathrm{Var}(X_i)=\sigma^2$. Then,*

$$\lim_{n\to\infty}\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}=\mathcal{N}(0,1).$$

There is an equivalent formulation: if $\overline{X}=\sum_{i=1}^n X_i/n$, then Theorem 19.3 states that $\overline{X}\sim\mathrm{N}\left(\mu,\sigma^2/n\right)$ as $n\to\infty$: if $Z=(\overline{X}-\mu)/\sqrt{\sigma^2/n}$, then $Z\sim\mathrm{N}(0,1)$, but also

$$Z=\frac{\sum_{i=1}^n X_i/n-\mu}{\sqrt{\sigma^2/n}}=\frac{n\left(\sum_{i=1}^n X_i/n-\mu\right)}{n\sqrt{\sigma^2/n}}=\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}.$$

This means that if one takes a large number of sample means of *any* distribution, they form a normal distirbution. This is why the normal distribution appears so often in the real world, and allows things such as election polling (a reasonable sample of people can lead to a probability that the whole group will vote one way). The distribution can be discontinuous, skewed, or anything else horrible. And this is why this theorem is so good.

However, polling can be misused: lots of election polls end up sampling nonrepresentative populations and wildly mis-predicting the election. Similarly, McDonalds rolled out the McRib, a flop, after successful testing in South Carolina, which clearly represents the rest of the nation in terms of barbecue.

Example 19.1. Applying this to the midterm, there are 230 X_i, and $E[X_i]=95.4$ and $\mathrm{Var}(X_i)=400.40$. Thus, one can create ten disjoint samples Y_i and their sample means $\overline{Y}_i$. Then, using the central limit theorem, these should all be turned into $Z\sim\mathrm{N}(0,1)$, which ended up being really close.

Example 19.2. Suppose one has an algorithm which has a mean running time of μ seconds and a variance of $\sigma^2=4$. Then, the algorithm can be run repeatedly (which makes for independent, identically distributed trials). Then, how many trials are needed in order to estimate $\mu\pm0.5$ with 95% certainty? By Theorem 19.3, if X_i is the running time of the i^{th} run, then

$$Z_n=\frac{\sum_{i=1}^n X_i-n\mu}{\sigma\sqrt{n}}=\frac{\sum_{i=1}^n X_i-n\mu}{2\sqrt{n}}\sim\mathrm{N}(0,1).$$

Thus,

$$\begin{aligned}P\left(-0.5\leq\frac{\sum_{i=1}^n X_i}{n-t}\leq0.5\right)&=P\left(\frac{-0.5\sqrt{n}}{2}\leq\frac{\sqrt{n}}{2}\frac{\sum_{i=1}^n X_i-nt}{n}\leq\frac{0.5\sqrt{n}}{2}\right)=P\left(\frac{-0.5\sqrt{n}}{2}\leq Z_n\leq\frac{0.5\sqrt{n}}{2}\right)\\&=\Phi\left(\frac{\sqrt{n}}{4}\right)-\Phi\left(-\frac{\sqrt{n}}{4}\right)=2\Phi\left(\frac{\sqrt{n}}{4}\right)-1\approx0.95,\end{aligned}$$

which means that (calculating Φ^{-1} with the table of normals) $\sqrt{n}/4\approx1.96$, or $n=\lceil(7.84)^2\rceil=62$.

Chebyshev's inequality provides a worse bound: $\mu_s=\mu$ but

$$\sigma_s^2=\mathrm{Var}\left(\sum_{i=1}^n\frac{X_i}{n}\right)=\sum_{i=1}^n\mathrm{Var}\left(\frac{X_i}{n}\right)=\frac{n\sigma^2}{n^2}=\frac{4}{n},$$

[30] Note that one still needs the probabilistic statement; things with probability zero can still happen, in some sense, but that requires a digression into measure theory.

so

$$P\left(\left|\sum_{i=1}^{n}\frac{X_i}{n}-t\right|\geq 0.5\right)\leq\frac{4/n}{0.5^2},$$

which implies that $n\geq 320$. This is a much worse bound.

Example 19.3. Suppose the number of visitors to a web site in a given minute is $X\sim\text{Poi}(100)$, and the server crashes if there are at least 120 requests in a given minute. Then, one can use the CLT to break this into lots of small distribiutions: $\text{Poi}(100)=\sum_{i=1}^{n}\text{Poi}(100/n)$, with all of the distributions independent, which allows one to use a normal distribution to approximate a Poisson one.

Example 19.4. Suppose someone rolls ten 6-sided dice, giving events $X_1,\ldots,X_{10}$. Let X be the total value of all of the dice. What is the probability that $X\leq 25$ or $X\geq 45$? Using the Central Limit Theorem, $\mu=E[X_i]=3.5$ and $\sigma^2=\text{Var}(X_i)=35/12$. Then, approximating with a normal distribution, one shows that there is about an 0.08 probability that this will happen. Note that Chebyshev would only say that there is at most a 0.292 probability of it happening.

20. Parameters: 5/15/13

All of the distributions specified so far are parametric models: the distribution is given in terms of parameters that yield the actual distribution, such as $\text{Uni}(\alpha,\beta)$, where α and β are the parameters. Generically, these parameters are denoted θ, so for a Poisson distribution $\text{Poi}(\lambda)$, $\theta=\lambda$, and sometimes it's a tuple: for a uniform dustribution, $\theta=(\alpha,\beta)$.

This is important because the true parameters aren't always known; instead, we only have experimental evidence. Thus, one might assume the real-world data is given by a distribution, and try to find that istribution.

Definition. An estimator is a random variable that represents a parameter.

If $X\sim\text{Ber}(p)$, p is the truth. If we can't see it, we might instead have a model $\hat{p}$ that estimates it, but then $\hat{p}$ is a random variable. The goal is to have a point estimate, which makes for a single value for the parameter rather than a distribution. This allows a better understanding of the process, yielding data, predictions, and simulations.

Suppose $X_1,\ldots,X_n$ are IID random variables. Then, the sample sample mean $\overline{X}=\sum X_i/n$ is a random variable, so one can take the sampling distribution of the mean, which is the distribution of $\overline{X}$. The Central Limit Theorem shows that this is a roughly normal distribution when n is large (specifically, it's a good approximation when $x>30$ and is very good when $x>100$). Then, we have $\text{Var}(\overline{X})=\sigma^2/n$, where $\text{Var}(X_i)=\sigma^2$, and we have $S^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$. For large n, the $100(1-\alpha)\%$ confidence interval is $(\overline{X}-z_{\alpha/2}S/\sqrt{n},\overline{X}+z_{\alpha/2}S/\sqrt{n})$, where $\Phi(z_{\alpha/2})=1-\alpha/2$. Here, $S/\sqrt{n}$ can be thought of as the square root of the standard variance. For example, if $\alpha=0.05$, then $\Phi(z_{\alpha/2})=0.975$, so $z_{\alpha/2}\approx 1.96$ with the help of a table of normals.

The meaning of a confidence interval is: with probability $1-\alpha$ when the confidence interval is computed from a sample, the true mean μ is in the interval. Note that this is *not* the probability that μ itself is in the interval for some given interval, which is either 0 or 1.

Example 20.1. Suppose a company wants to estimate the average number of hours a CPU spends idle. Then, 225 computers are monitored. Suppose $\overline{X}=11.6$, $S^2=16.81$ hours, and $S=4.1$ hours. Thus, for a 90% confidence interval, $z_{\alpha/2}=1.645$ (which is another common quantity you'll see), so the interval is

$$\left(\overline{X}-z_{\alpha/2}\frac{S}{\sqrt{n}},\overline{X}+z_{\alpha/2}\frac{S}{\sqrt{n}}\right)=\left(11.6-1.645\frac{4.1}{\sqrt{225}},1.6+1.645\frac{4.1}{\sqrt{225}}\right)=(11.15,12.05).$$

Suppose $X_1,\ldots,X_n$ is a set of IID random variables. Then, the empirical n^{th} moments are $\hat{m}_1=\sum_{i=1}^{n}X_i/n$. These are also referred to as the sample moments, and $\hat{m}_1$ is the sample mean. The method of moments uses the approximation $m_i\approx\hat{m}_i$ to estimate the parameters of the model. For example,

$$\text{Var}(X)\approx\hat{m}_2-(\hat{m}_1)^2=\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-\overline{X}^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n}\sum_{i=1}^{n}\overline{X}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i^2-\overline{X}^2).$$

Notice that this is different than the sample variance, which is $S^2=n(\hat{m}_2-\hat{m}_1^2)/(n+1)$. This difference is significant when n is small, so consider a sample of size 1. Then, the sample variance is undefined, because the variability is meaningless with only one data point. However, the method of moments just reports the variance as zero, since only the one data point matters.

The bias of an estimator $\hat{\theta}$ is $E[\hat{\theta}]-\theta$. An estimator with a bias of zero is called unbiased.[31] For example, the sample mean is unbiased: $E[\overline{X}]=\mu$. Then, the sample variance is unbiased, but the method of moments is biased (though as $n\to\infty$ it becomes less asymptotically biased).

[31] Bias is a loaded word in the outside world, but depending on the application having some bias in an estimator may cause more accurate results.

An estimator $\hat{\theta}$ is (weakly) consistent if for any $\varepsilon > 0$, $\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$. This intuitively means that as more and more data is fed to the estimate, it gets more and more accurate. Note the similarity in the expression to that of Proposition 19.1. The sample mean is also consistent, but there is not generally a relation between consistency and presence or lack of bias. Method-of-moments estimates are also generally consistent.

Example 20.2. Suppose $X_1, \ldots, X_n$ are IID random variables, with $X_i \sim \text{Poi}(\lambda)$. Then,

$$\lambda = E[X_i] \approx \hat{m}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X - i = \hat{\lambda}.$$

Since $\lambda = \text{Var}(X_i)$ as well, one has another estimate of λ which isn't always the same:

$$\lambda = E[X_i^2] - E[X_i]^2 \approx \hat{m}_2 - \hat{m}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \overline{X}^2) = \hat{\lambda}'.$$

In general, the estimator of the lowest order or the one that is easiest to compute is used, and they tend to be the same.

Example 20.3. If $X_1, \ldots, X_n$ are IID and $X_i \sim \mathsf{N}(\mu, \sigma^2)$, then $\mu \approx \hat{\mu} = \sum X_i/n$, but estimated variance is not equal to the sample distribution. One is biased, and the other isn't. If $X_i \sim \text{Uni}(\alpha, \beta)$, one can obtain similar expected means $\hat{\mu}$ and variance $\hat{\sigma}^2$, which leads to a system of equations in α and β: $\hat{\alpha} = \overline{X} - \hat{\sigma}\sqrt{3}$ and $\beta = \overline{X} + \hat{\sigma}\sqrt{3}$. Notice that there could be data outside of these data points, so be careful.

21. Likelihood: 5/17/13

Definition. Consider n IID variables $X_1, \ldots, X_n$. In some sense, X_i is a sample from a density function $f(X_i \mid \theta)$ for some parameter θ. Then, the likelihood function is $L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$.

This represents how likely some set of data $(x_1, \ldots, x_n)$ is given the density, though it becomes a product because the X_i are indepedent.

Definition. The maximum likelihood estimator (MLE) of θ is the value (or set of values, if θ is a vector) of θ that maximizes the likelihood function $L(\theta)$. More formally, $\theta_{\mathsf{MLE}} = \arg\max_\theta L(\theta)$.[32]

For this, it is often more convenient to use the log-likelihood function

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i \mid \theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta).$$

Here, log denotes the natural log, though the base doesn't matter for maximizing. This still works for maximizing θ because it is monotone, so $x \le y$ iff $\log x \le \log y$ for $x, y > 0$. Thus, $\arg\max_\theta L(\theta) = \arg\max_\theta LL(\theta) = \arg\max_\theta cLL(\theta)$ for any positive constant c, and this can be found with a bit of linear algebra:

- Find the formula for $LL(\theta)$ and differentiate it with each $\frac{\partial LL(\theta)}{\partial \theta}$, and set it to zero.
- This is a set of equations, which can be solved simultaneously.
- Then, it's necessary to check that the critical points aren't minima or saddle points by checking nearby values.

Example 21.1. Suppose $X_1, \ldots, X_n$ are IID random variables such that $X_i \sim \text{Ber}(p)$. Then, the PMF can be rewritten in the form $f(X_i \mid p) = p^{x_i}(1-p)^{1-x_i}$, where $x_i = 0$ or $x_i = 1$. This has the same meaning, but is now differentiable. Thus, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i},$$

and therefore the log-likelihood is

$$LL(\theta) = \sum_{i=1}^{n} \log(p^{X_i}(1-p)^{1-X_i}) = \sum_{i=1}^{n} (X_i \log p + (1 - X_i)\log(1-p)) = Y \log p + (n - Y)\log(1-p),$$

where $Y = \sum_{i=1}^{n} X_i$. After a bunch of calculus, the MLE is $p_{\mathsf{MLE}} = Y/n = \sum_{i=1}^{n} X_i$. This agrees with intuition, and looks like sample mean and the method of moments for this distribution.

[32] The function $\max(f(x))$ returns the maximum of $f(x)$, but $\arg\max(f(x))$ returns the value x such that $f(x)$ is maximized.

Example 21.2. Suppose the $X_1, \dots, X_n$ are now IID Poisson variables, $X_i \sim \text{Poi}(\lambda)$. Then, the PMF is $f(X_i \mid \lambda) = e^{-\lambda}\lambda^{x_i}/x_i!$, which makes the likelihood $\prod e^{-\lambda}\lambda^{x_i}/x_i!$. Thus, the log-likelihood is

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n}(-\lambda \log e + X_i \log\lambda - \log(X_i!)) \\ &= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\log(X_i!). \end{aligned}$$

This looks alarming until you realize we only have to differentiate with respect to λ, showing that $\lambda_{\text{MLE}} = \sum X_i/n$.

Example 21.3. If $X_i \sim \text{N}(\mu, \sigma^2)$, then one obtains the log-likelihood

$$L(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-(X_i-\mu)^2/(2\sigma)^2}}{\sigma\sqrt{2\pi}}\right) = \sum_{i=1}^{n}\left(\log\left(\sigma\sqrt{2\pi}\right) - \frac{(X_i-\mu)^2}{(2\sigma^2)}\right).$$

Now, there are two things to differentiate:

$$\frac{\partial LL(\mu,\sigma^2)}{\partial\mu} = \sum_{i=1}^{n}\frac{2(X_i-\mu)}{(2\sigma^2)} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0,$$

and σ (not σ^2, since this is with respect to a variable):

$$\frac{\partial LL(\mu,\sigma^2)}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^3} = 0.$$

Then, $\sum X_i = n\mu$, so $\mu_{\text{MLE}} = \sum X_i/n$, which is the sample mean again. This is useful, but there are some distributions where it isn't the sample mean. Notice that μ is unbiased, but σ is biased:

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu^2_{\text{MLE}}).$$

This seems counterintuitive, but there's no reason that it shouldn't be the case.

Example 21.4. Now, suppose $X_i \sim \text{Uni}(\alpha, \beta)$. Then, the PDF is $f(X_i \mid \alpha, \beta) = 1/(\beta - \alpha)$ if $\alpha \le x_i \le \beta$ and is zero otherwise, so the overall likelihood is

$$L(\theta) = \begin{cases} \left(\frac{1}{\beta-\alpha}\right)^n, & \alpha \le x_1, \dots, x_n \le \beta \\ 0, & \text{otherwise.} \end{cases}$$

This can be solved with Lagrange multipliers, which is complicated and unfortunate, so turn instead to intuition. The goal is to make the interval as small as possible while guaranteeing that all of the values are still within the region, which means that $\alpha_{\text{MLE}} = \min(x_1, \dots, x_n)$ and $\beta_{\text{MLE}} = \max(x_1, \dots, x_n)$.

These do interesting things in small sample sizes. In many cases, the sample mean is obtained, which is pretty nice. But the normal distribution is biased, and underestimates for small n, and the uniform distribution is also biased, and behaves badly for small samples (e.g. if $n = 1$, so that $\alpha_{\text{MLE}} = \beta_{\text{MLE}}$, which isn't necessarily right). This just fits the distribution best to what has already been seen, but this doesn't always generalize well and the variance might be too small.

However, the MLE is frequently used, because it has lots of nice properties:

- It's consistent: $\lim_{n\to\infty} P(|\theta_{\text{MLE}} - \theta| < \varepsilon) = 1$ for any $\varepsilon > 0$.
- Though it is potentially biased, the bias disappears asymptotically as n increases.
- The MLE has the smallest estimate of all good estimators for large samples.

This, this is frequently used in practice where the sample space is large relative to the parameter space. However, this is a problem if the variables are dependent: joint distributions cause the number of parameters of n variables to be about 2^n. This can cause issues in modeling.

Example 21.5. Suppose $Y_1, \dots, Y_n$ are independently and identically distributed random variables, such that $Y_i \sim$ Multinomial$(p_1, \dots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$, and let X_i be the number of trials of outcome i where $\sum_{i=1}^{m} X_i = n$. Thus, the PDF is

$$f(X_1, \dots, X_m \mid p_1, \dots, p_m) = n!\prod_{i=1}^{m}\frac{p_i^{x_i}}{x_i!}.$$

Thus, the log-likelihood is

$$LL(\theta) = \log n! - \sum_{i=1}^{m} \log(X_i!) + \sum_{i=1}^{m} X_i \log p_i.$$

This time, Lagrange multipliers are necessary. The goal is to maximize

$$A(\theta) = \sum_{i=1}^{m} X_i \log p_i + \lambda \sum_{i=1}^{m} p_i - 1,$$

so after differentiating, $p_i = -X_i/\lambda$, so using the initial constraints $\sum X_i = n$ and $\sum p_i = 1$ one can see that $\lambda = -n$. Intuitively, this means that the probability p_i is the proportion of outcome i. This makes sense.

If one has a six-sided die and rolls 12 rolls, none of which is a 3, which suggests that maybe one would never roll a 3. In the frequentist interpretation (probability is frequency in the limit). more rolls are necessary to understand this, but there's another interpretation. Bayesian probability is a belief in something, maybe given some prior information, but still a belief. If you were asked the probability of it raining tomorrow, how many tomorrows have you experienced in order to make a valid prediction?

Example 21.6. Suppose one has two envelopes, one of which contains $\$X$ and the other of which contains $\$2X$. If someone takes one envelope and opens it, seeing $\$Y$, the expected value of switching the envelope is $5Y/4$. But this didnt depend on knowing the value of Y, so one should always switch envelopes... even after one has already switched... right?

22. More Bayesian Probability: 5/20/13

Returning to Example 21.6, we know that if Y is the money in the envelope one selects and Z is the money in the other envelope, then $E[Z \mid Y] = 5Y/4$, which leads to the confusing switching strategy. Various sorts of bringing other information in (e.g. if there's an odd number of cents in one) can lead to better guesses. However, the switching strategy assumes that all values of X (the smaller amount of money in the envelopes) are equally likely, but this means that the PDF would have to be everywhere zero.

This is an example of Bayesian probability, which is more subjective: since the distribution can't be uniform and it must exist, then what is it? The frequentist must play infinitely many times to determine this, but if only one trial is allowed, the Bayesian must work based on prior beliefs: what would be the most reasonable value to see here? This leads to the prior belief of the distribution of X.

For example, suppose someone has a coin that's never been flipped before. It seems reasonable that it will be a fair coin, but do we know that? A purely frequentist approach has no idea what the answer is.

A Bayesian would approach the envelope problem by determining a prior value for $E[Z \mid Y]$, and then keep the envelope if $Y > E[Z \mid Y]$, and switch otherwise. Opening the envelope provides data to compute $P(X \mid Y)$ and therefore $E[Z \mid Y]$. In order to make any decision at all, one has to make a prior decision.[33]

In this interpretation, Bayes' theorem $P(\theta \mid D) = P(D \mid \theta)P(\theta)/P(D)$ can be thought of with θ as the parameters of a model and D as the data. Then, $P(\theta)$ is the prior estimate, and $P(D \mid \theta)$ is the likelihood this has been seen before. Then, the updated (posterior) probability is $P(\theta \mid D)$. However, $P(D)$ is more nuanced to calculate: it can be found by $P(D) = \int P(D \mid \theta)P(\theta)\,\mathrm{d}\theta$, but it is better to think of as a constant that turns the left-hand side into a probability distribution.

Example 22.1. Suppose one has a prior distribution $\theta \sim \text{Beta}(a, b)$ with data D that is n heads and m tails. Then,

$$\begin{aligned} f_{\theta|D}(\theta = p \mid D) &= \frac{f_{D|\theta}(D \mid \theta = p) f_\theta(\theta = p)}{f_D(D)} \\ &= \frac{\binom{n+m}{n} p^n (1-p)^m (p^{a-1}(1-p)^{b-1})}{C_1 C_2} = C_3 p^{n+a-1}(1-p)^{m+b-1}. \end{aligned}$$

Here, C_1 and C_2 are the original normalizing constants for the original distributions, and C_3 is the normalizing distribution for the final distribution. θ is a prior distribution based on what one thinks the data is before any trials. These can be very different, but as long as enough trials are done, the posteriors will always converge to the true value.

[33] Frequentist and Bayesian viewpoints are the closest thing probability has to a religion. The professor, for example, is a Bayesian, and the author of the textbook is very frequentist.

Recall the MLE $\theta_{\text{MLE}} = \arg\max_\theta \prod_{i=1}^n f(X_i \mid \theta)$; one has a more real-world estimate of θ given the data. This is known as the maximum *a posteriori* estimate: suppose $X_1, \ldots, X_n$ are IID.

$$\theta_{\text{MAP}} = \arg\max_\theta f(\theta \mid X_1, \ldots, X_n) = \arg\max_\theta \frac{f(X_1, \ldots, X_n \mid \theta) g(\theta)}{h(X_1, \ldots, X_n)}$$
$$= \arg\max_\theta \frac{\left(\prod_{i=1}^n f(X_i \mid \theta)\right) g(\theta)}{h(X_1, \ldots, X_n)|} = \arg\max_\theta g(\theta) \prod_{i=1}^n f(X_i \mid \theta),$$

where $g(\theta)$ is the prior distribution. Notice that there is a significant philosophical distinction, but not as much of a mathematical distribution. Again, it can be convenient to use logarithms:

$$\theta_{\text{MAP}} = \arg\max_\theta \left(\log g(\theta) + \sum_{i=1}^n \log(f(X_i \mid \theta)) \right).$$

Recall the beta distribution with parameters a and b. These are called hyperparameters, because they are parameters that determine the parameters of the distribution. For example, a fair coin is given a prior Beta(x, x) distribution, with higher values of x correlated with a lower variance.

In order to do this with a multinomial distribution, introduce the Dirichlet distribution: the way that the beta distribution is conjugate to the binomial (or Bernoulli) distribution is the way the Dirichlet distribution is conjugate to the multinomial distribution. This distribution is written Dirichlet$(x_1, \ldots, x_n)$:

$$f(x_1, \ldots, x_n) = \frac{1}{B(a_1, \ldots, a_m)} \prod_{i=1}^m x_i^{a_i - 1}.$$

The prior (imaginary trials) represent seeing $\sum_{i=1}^m a_i - m$ imaginary trials such that $a_i - 1$ are of outcome i. Then, after observing $n_1 + \cdots + n_m$ new trials (such that n_i had outcome i), the new distribution is Dirichlet$(a_1 + n_1, \ldots, a_m + n_m)$. Here is an animation of the logarithmic density of Dirichlet(a, a, a).

The Dirichlet model allows one to work around the multinomial problem presented in the previous lecture: if one rolls a die but never sees a 3, one could do an MAP estimate in which some number of prior trials did give a 3. The MAP estimate for Dirichlet$(k + 1, \ldots, k + 1)$ is $p_i = (X_i + k)/(n + mk)$ for this experiment. Now, the probability of rolling a 3 is nonzero, which is nice. This method with $k = 1$ is also known as Laplace's law. Notice that this is a biased estimate.

The conjugate distribution for the Poisson and exponential distributions is Gamma(α, λ), which represents a prior of α imaginary events in a time period of λ. To update this for a posterior distribution, one observes n events in k time periods, leading to a posterior distribution of Gamma$(\alpha + n, \lambda + k)$.

Finally, conjugate to the normal distribution there are several of these depending on which of the mean and variance are known. For example, Normal(μ_0, σ_0^2) where σ^2 is known but μ isn't, is a hyperparameter, where one believes the true $\mu \sim \mathsf{N}(\mu_0, \sigma_0^2)$. After observing n data points, the posterior distribution is

$$\mathsf{N}\left(\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right), \left(\frac{1}{\sigma^2} + \frac{n}{\sigma^2} \right)^{-1} \right).$$

Generally, one doesn't know the variance without other information, but it could be estimated without knowing the mean, making this useful.

23. Machine Learning: 5/22/13

Today, the parameter estimation that has been discussed in the previous lectures will be grounded with some actual algorithms. These methods will be applied in the next (and last) problem set.

Notice that in the real world, one doesn't have perfect distributions; instead, there's just data, and one should try to make predictions based on the data. Often, with enough data, one doesn't need any expertise in the specific field, since it can be viewed as abstraction.

Machine learning in general is pretty useful, and the small amount covered in this class will form a sample of the AI that every CS major should know. Machine learning is a wide field, but this class will deal with a subfield of preduction: given some data, one wishes to build a model for it that allows for the making of predictions. The probability formalism is very useful for this process.

Formally, one has a vector of m observed variables $\mathbf{X} = \langle X_1, \ldots, X_m \rangle$. These are called the input features or input variables. They are the things we can observe about the world, and are fed into the black-box algorithm.[34] Then, the

[34]Sometimes statisticians call these independent variables (as opposed to dependent variables), but they are often not independent in the probabilistic sense. Be careful.

goal is to make a prediction for a variable that we don't know, called the output feature or variable Y. There are generalizations in which Y is a vector, but they won't be considered here. Y is called a dependent variable.

The goal is to "learn" a prediction function $g(\mathbf{X})$ that spits out predictions of Y: $\hat{Y} = g(\mathbf{X})$. When Y is discrete, this is called a classification task, and when Y is continuous, this is called a regression.

Machine learning has lots and lots of applications:

- Stock prediction: one has a lot of economic variables (e.g. mortgage rates, interest rates from the Fed, current stock prices), and wishes to predict what the stock market does. If a sufficiently good model can be found, then there's a nice way to make money (e.g. D.E. Shaw).
- Computational biology: given some input data (one's DNA and environmental risk factors), what is the likelihood for getting a certain disease?
- This can be used to predict what products a set of consumers might prefer.
- Credit card fraud can be detected: a model can be set up for a given user using machine learning, and actions particularly outside of that model is flagged. For a while, AT&T was a world leader in machine learning because of credit and telephone fraud.
- Spam detection is another example. In the header of some email systems, there's a field called the Bayesian spam probability, which indicates a prediction as to how spammy an email might be given the distribution of the words and a model. The order of the words is completely ignored, interestingly. Using increasingly sophisticated models, more and more spam is correctly detected even as more and more email sent is spam, illustrating that models are generally better when they have more data.

The data is given as a set of N training instances, each of which is a data points (often a vector of values). This is a pair $(\langle x_1, \dots, x_m \rangle, y)$ of observed inputs and the correct result given those inputs. These are previously observed data, or can be thought of as historical data. This illustrates what we want to predict, and with enough data, the probabilistic dependencies can be elucidated.

The predictor g is often parametric: it has some parameters that get determined by the training data, such as $g(X) = aX + b$, in which the parameters are a and b. The choice of parameters is something that can be discussed in considerably more detail; take CS 229 if interested. For a regression, the goal is to minimize the mean-squared error $E[(Y - g(\mathbf{X}))^2]$ (MSE), which is sometimes known as a loss function (in which the lost object is truth). There are other ways to judge models, such as mean absolute value or logarithmic ones; the former is suboptimal because it isn't differentiable at zero. For classification, the best Y has generally $g(\mathbf{X}) = \arg\max \hat{P}(Y \mid \mathbf{X})$. Under some surprisingly mild assumptions, this is the best that can be done.

Consider a set of N training instances $(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^n, y^n)$. (Here, superscripts are used as indices, not powers). The learning algorithm processes these inputs into a classifier g, which can be used to make predictions for new datasets.

Example 23.1. One concrete example is linear regression in which a variable Y is predicted given a single variable X. One assumes the model is linear: $\hat{Y} = g(X) = aX + b$. Then, given N pairs of points $(x^1, y^1), \dots, (x^n, y^n)$. Geometrically, one has a set of points in the plane, and which line minimizes the error of the predictions? Typically, the mean-squared error is used.

Some math shows that the mean-squared error is $E[(Y - aX - b)^2]$, and its partial derivatives with respect to a and b can be calculated:

$$\frac{\partial}{\partial a} E[(Y - aX - b)^2] = E[-2X(Y - aX - b)] = -2E[XY] + 2aE[X^2] + 2bE[X]$$
$$\frac{\partial}{\partial b} E[(Y - aX - b)^2] = E[-2(Y - aX - b)] = -2E[Y] + 2aE[X] + 2b.$$

This gives a system of equations whose solution is $a = \operatorname{Cov}(X, Y)/\operatorname{Var}(X) = \sigma_y/\sigma_x \rho(x, y)$ and $b = \sigma_y/\sigma_x \rho(x, y)\mu$. These quantities are obtained as estimates from the training data, using any of the methods discussed in previous lectures.

Example 23.2. Now, suppose X and Y are both discrete, making this a classification problem: suppose $X \in \{1, 2, 3, 4\}$. One way this can be done is to take a continuous range and pick several subsets (e.g. the temperature is in the 60s, 70s, or 80s, which form the discrete domain on three elements). Then, Y is the weather for the next day: either rainy or sunny the next day.

Given the training data, one can estimate the joint PMF $\hat{p}_{X,Y}(X, Y)$, and then, using Bayes' theorem, one has $P(Y \mid X) = p_{X,Y}(x, y)/p_X(x)$. Then, for a new X, predict $\hat{Y} = g(\mathbf{X}) = \arg\max_Y \hat{P}(Y \mid \mathbf{X})$. However, since $p_X(x)$ isn't modified by the choice of Y, it can be ignored (since one takes the arg max anyways): $\hat{Y} = \arg\max_Y \hat{P}(X \mid Y)\hat{P}(Y)$.

There is more than one way to estimate the joint PMF:

- The MLE method just counts the number of times each pair appears and then normalizes them to get probability.

- One can also add a Laplace prior: add 1 to each pair (not just the ones that appear in the training data), and normalize so that the total value is 1. This is useful if all of the events have some nonzero probability and is thus helpful when the data is limited.

A new observation of $X = 4$ can be used to make a prediction of the weather tomorrow: as $\hat{Y} = \arg\max_Y \hat{P}(X, Y) = \arg\max_Y \hat{P}(X \mid Y)\hat{P}(Y)$. Here, the MLE and Laplace estimates agree (with the given table). The similarities and differences are worth noting. Specifically, the probabilities differ: without the Laplace prior, some events seem to have zero probability, which might not happen in the real world. It is a biased estimate in that we introduce our own subjective belief about the world, but this is sometimes better.

Now, suppose there are m input values $\mathbf{X} = \langle X_1, \dots, X_m \rangle$. In theory, one could just take the table of $P(X_1, \dots, X_n \mid Y)$ for all possible values of all of the X_i, which is incredibly huge and therefore not all that insightful: most of the entries will be zeros, so the predictions won't be all that helpful. The asymptotic size is actually important: it's exponential in m, which was a problem in simulations in which $m \approx 10000$.

Thus, a simpler model is necessary: enter the Naïve Bayesian Classifier[35] This makes the assumption that all of the X_i are conditionally independent given Y. This is often untrue in practice, but allows the predictions to be simplified while still being reasonably accurate.

The beauty of this is that the conditional probability factors:

$$P(\mathbf{X} \mid Y) = P(X_1, \dots, X_m \mid Y) = \prod_{i=1}^{m} P(X_i \mid Y),$$

so the table now has linear space, with only the one probabilistic assumption!

Example 23.3. Let X_1 be an indicator variable that indicates that someone likes Star Wars, and X_2 is that a person likes Harry Potter. Then, Y is an indicator variable for whether someone likes the Lord of the Rings. This is a standard marketing application, for finding recommended products, and previous purchases can be used to make a model. Thus, we have the joint and marginal probability tables.

Now, suppose a person comes along and $X_1 = 1$, but $X_2 = 0$. Then, under the Naïve Bayes assumption, $\hat{Y} = \arg\max_Y \hat{P}(\mathbf{X} \mid Y)\hat{P}(Y) = \arg\max_Y \hat{P}(X_1 \mid Y)\hat{P}(X_2 \mid Y)\hat{P}(Y)$. The arg max can be evaluated by picking $Y = 0$ and $Y = 1$ and picking whichever makes the result better.

The number of parameters can grow very large: in one model of spam classification, $m \approx 100000$, and there are m variables corresponding to indicators for whether a given word (out of *all* English words) appeared in an email. Since m is extremely large, then the Naïve Bayes assumption is necessary. Additionally, since there are lots of words that could be used in spam but that aren't in the sample dataset, the Laplace estimate is probably a good idea. Then, the data is tested with some other testing data (you can't use the same training data; that would be cheating).

There are various criteria for effectiveness, such as precision (correctly predicted the most cases relative to incorrect ones) or recall (correctly predicted the most cases relative to all cases).

24. Logistic Regression: 5/24/13

There are a couple formulas that might be helpful for the next problem set, regarding a Naive Bayesian classifier: $\hat{P}(Y = 0)$is the number of instances in the class with value 0 over the total number of instances, and $\hat{P}(X_i = 0, Y = 0)$ is the number of instances where $X_i = 0$ and the class is zero. Thus, $\hat{P}(X_i = 0 \mid Y = 0) = \hat{P}(X_i = 0, Y = 0)/P(Y = 0)$ and $\hat{P}(X_i = 0 \mid Y = 1) = \hat{P}(X_i = 0, Y = 1)/\hat{P}(Y = 1)$. Finally, $P(X_i = 1 \mid Y = 0) = 1 - \hat{P}(X_i = 0 \mid Y = 0)$, and the estimate for y is $\hat{y} - \arg\max_y P(\mathbf{X} \mid Y)P(Y) = \arg\max_y(\log(P(\mathbf{X} \mid Y)P(Y)))$. The log estimate is helpful to prevent numerical underflow (i.e. small data values becoming eaten by rounding error). It can be computed specifically as

$$\log P(\mathbf{X} \mid Y) = \log(P(X_1, \dots, X_m \mid Y)) = \log \prod_{i=1}^{m} P(X_i \mid Y) = \sum_{i=1}^{m} \log P(X_i \mid Y).$$

This is the MLE formula, and the formula for the Laplacian is similar. In some sense, this classifier models the joint probability $P(\mathbf{X}, Y)$. However, the goal is to find $P(Y \mid \mathbf{X})$ (since $\mathbf{X}$ was observed, and Y is to be predicted), so the goal is to find $\hat{y} = \arg\max_y P(Y \mid \mathbf{X})$.

One can use a technique called logistical regression to model $P(Y \mid \mathbf{X})$, which can be called the conditional likelihood. This uses a function called the logistic function (also sigmoid or "squashed S" function): $P(Y = 1 \mid \mathbf{X}) = 1/(1 + e^{-z})$,[36] where $z = \alpha + \sum_{j=1}^{m} \beta_j X_j$. In some sense, one takes this nice linear function and runs it through the logistic function.

[35] This is apparently referred to as the Idiot Bayes Classifier by statisticians.

[36] This is an arctangent function, strangely enough.

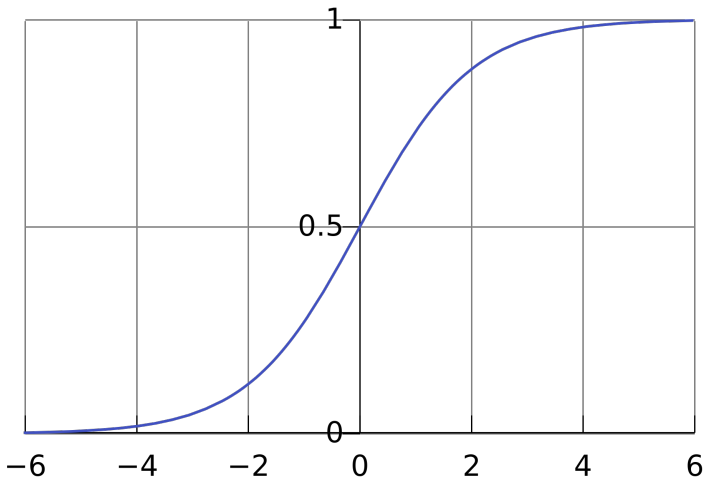


FIGURE 7. The standard logistic function. Source

The parameters that can change are the α and β_j, and often, the simplifying trick of letting $X_0 = 1$ and $\beta_0 = \alpha$, so the sum becomes $z = \sum_{j=0}^{m} \beta_j X_j$.

Since $Y = 0$ or $Y = 1$, then $P(Y = 0 \mid \mathbf{X}) = e^{-z}/(1 + e^{-z})$, with z as before. This still doesn't illuminate why the function was chosen, but calculate the log-odds:

$$\log \frac{P(Y = 1 \mid \mathbf{X})}{P(Y = 0 \mid \mathbf{X})} = \log \frac{1}{e^{-z}} = z.$$

This functional form guarantees that the log-odds is a linear function, which is pretty convenient.

Now, the log-conditional likelihood of the data is

$$LL(\theta) = \sum_{i=1}^{n} (y_i \log(P(Y = 1 \mid \mathbf{X})) + (1 - y_i) \log(P(Y = 0 \mid \mathbf{X}))).$$

This sum represents dividing the data into two cases, those with $y_i = 0$ and $y_i = 1$. Given this, one might want to know the β_j that create the MLE, but, unfortunately, there is no analytic derivation of it.

However, the log-conditional likelihood function is concave, so it must have a single global maximum. This can be repeatedly estimated by computing a gradient repeatedly somewhat like Newton's Method:

$$\frac{\partial LL(\theta)}{\partial \beta_j} = x_j \left(y - \frac{1}{1 - e^{-z}} \right) = x_j (y - P(Y = 1 \mid \mathbf{X})).$$

Thus, $LL(\theta)$ can be iteratively updated using the formula

$$\beta_j^{\text{new}} = \beta_j^{\text{old}} + c x_j \left(y - \frac{1}{1 - e^{-z}} \right),$$

where $z = \sum_{j=0}^{m} \beta_j^{\text{old}}$ and c is a constant representing the step size.[37] Then, this is done repeatedly, until a given stopping criterion is met: one may just run it a fixed number of times or until some error bound is met.

[37] One should do this with care, since if c is too large, it might diverge, and smaller step sizes will take longer to converge. There are several ways to choose c, and one of them is called annealing, in which c starts large and gradually decreases.

Here's the derivation:

$$\begin{aligned} LL_i(\theta) &= y_i \log P(Y=1 \mid \mathbf{X}) + (1-y_i)P(Y=0 \mid \mathbf{X}) \\ &= y_i \log \frac{P(Y=1 \mid \mathbf{X})}{P(Y=0 \mid \mathbf{X})} + \log P(Y=0 \mid \mathbf{X}) \\ &= y_i \sum_{j=0}^{m} \beta_j X_j + \log e^{-z} - \log(1+e^{-z}) \\ &= y_i \sum_{j=0}^{m} \beta_j X_j - \sum_{j=0}^{m} \beta_j X_j - \log(1+e^{-z}). \\ \implies \frac{\partial LL_i(\theta)}{\partial \beta_j} &= yx_j - x_j + \frac{x_j e^{-z}}{1+e^{-z}} = x_j \left(y - \frac{1+e^{-z}}{1+e^{-z}} + \frac{e^{-z}}{1+e^{-z}} \right) \\ &= x_j \left(y - \frac{1}{1+e^{-z}} \right). \end{aligned}$$

Here, $y = y_i$, though it's not actually ambiguous because the function is $LL_i(\theta)$ only. However, the above model seems to only compute the gradient based on one data point. which seems suboptimal. This isn't actually the case, since the contributions to the gradient by each data point are all averaged to give the next step.

This is known as a batch logistic regression algorithm, and is an example of a gradient-ascent algorithm:

(1) Initialize $\beta_j = 0$ for all $0 \leq j \leq m$. There's no particular reason to initialize them in a specific place, so zero is a good choice.
(2) An epoch is the number of passes over the data that the algorithm makes. Thus, for each `i` less than the number of epochs,
    (a) Initialize the gradient vector `gradient` to zero.
    (b) Then, for each instance $(\mathtt{x}_1, \ldots, \mathtt{x}_m, y)$ in the training data:
        (i) For each $\mathtt{x_j}$, update the gradient:

$$\texttt{gradient[j]} += x_{\mathtt{j}} \left(y - \frac{1}{1+e^{-z}} \right),$$

    where $z = \sum_{\mathtt{j}=0}^{m} \beta_{\mathtt{j}} x_{\mathtt{j}}$. This is the batch part of the algorithm.
(3) Finally, update all of the $\beta_{\mathtt{j}}$ as $\beta_{\mathtt{j}} += \eta \cdot \texttt{gradient[j]}$ for $0 \leq \mathtt{j} \leq m$. Here η is a preset constant called the learning rate.

This determining of the parameters β_j is called training. Once this has been done, one can test the classifier: for each new test instance $\mathbf{X}$, compute $p = P(Y=1 \mid \mathbf{X}) = 1/(1+e^{-z})$, where $z = \sum_{j=0}^{m} \beta_j X_j$ as before, and classify $\hat{y} = 1$ if $y > 0.5$ and 0 otherwise. Notice that the β_j are not updated during the testing phase, though in some real systems, they may be updated periodically.

Geometrically, the goal of a logistic regression is to find a hyperplane in a high-dimensional space that separates the data instances where $Y = 1$ from those where $Y = 0$ (i.e. it is $(n-1)$-dimensional where the space is n-dimensional; the key is that it's linear). Functions or data sets in which this can be done are called linearly separable. Naïve Bayes is also a linear function, and even has the same functional form, but they compute different things: Naïve Bayes tries to best fit $P(\mathbf{X}, Y)$, and the logistic regression models $P(Y \mid \mathbf{X})$, though both use the same linearity assumption. Thus, their solutions are given by different algorithms.

Not all functions are linearly separable; just take the `xor` function. In practice, though, data that isn't linearly separable can still be approximated reasonably well. This dilemma halted research in neural networks for a couple of decades.

Notice that Naïve Bayes computes the joint probability in order to calculate the conditional probability, which is a bit of extra work (though it could use $P(\mathbf{X}, Y)$ to generate new data points). The logistic regression model is a discriminative model, because it doesn't model the whole distribution, but just tries to separate the data.

Logistic regression can be generalized when Y takes on more than two values, in which case the vector of β_j becomes a matrix. There are other choices that one may wish to make:

- What if the input variables are continuous, rather than discrete? The logistic regression is fine, though Naïve Bayes has a harder time. Specifically, some parametric form is neessary, or one could discretize the continuous values into ranges.
- If the values are discrete, Naïve Bayes handles it fine, because multi-valued data is just a multinomial distribution. The logistic regression has a harder time, since it requires its inputs to be numbers (unlike the Naïve Bayes model), which sometimes misrepresents the data.

The overall goal of machine learning is to build new models that generalize well to predicting new data. Sometimes, this even causes something called overfitting, in which the model fits so well that generality is lost. For example, one could use am interpolating polynomial to exactly hit all data points when a linear approximation is more likely to be correct (small errors in the data may be amplified in overfitting). The logistic regression is more prone to overfitting, because it doesn't model the whole distribution, but only focuses on Y. It can also happen if the training and testing data differ significantly (so that the nuances of the training data are overstated, as in McDonald's McRib release discussed earlier in this class). There are methods to mitigate overfitting in logistic regression, analogous to using priors in Naïve Bayes, though the optimization proces sis more complicated because there isn't an analytic solution.

In some sense, a logistic regression takes n inputs x_i with parameters β_i is a one-node neural network. Thus, one could connect many of them to create a neural network, though the algorithms are trickier: something called back-propagation allows one to approximate any function with a sufficient number of nodes, something called a uniform approximator.

25. Bayesian Networks: 5/29/13

With regards to machine learning, it is important to keep some perspective: there are these powerful tools for making predictions, but viewing these algorithms as black boxes isn't ideal. It is generally also useful to understand the domain: what things are correlated, what things are independent, etc.

For example, palm size negatively correlates with life expectancy. This surprising result occurs because women tend to have smaller palms than men, but on average live longer. Here, the correlation is definitely not the whole story, in a reasonably obvious way — but what about all of the other medical correlations that might be explained by other effects?

The way to formalize this is something called a Bayesian network (for more on this subject, take CS 228 or 228T). This is a directed graph in which the nodes are random variables, and an arc from X to Y is present if X has a direct influence on Y, in which case X is called a "parent" of Y.[38] Then, each node X has probability $P(X \mid \text{parents}(X))$. Finally, in order for the math to make sense, it is necessary that there are no cycles (where direction matters), so that the graph is a directed acyclic graph (DAG).

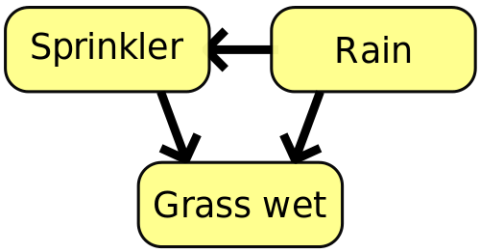


Figure 8. An example of a simple Bayesian network. Source

One important idea is that each node is conditionally independent of its non-descendants, given the parents. For example, given gender G, palm size S and life expectancy L are independent: $P(S, L \mid G) = P(S \mid G)P(L \mid G)$. With each node one can associate a conditional probability table (CPT) of the probability of it taking on some value given the parents.

Thus, it is possible to obtain all of the joint probabilities given the CPT: the conditional independence of the graph modularizes the calculation, yielding

$$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(X_i \mid \text{parents}(X_i)).$$

The CPT is useful because the number of parameters in a full table is exponential in the number of nodes, but with this setup, it's linear, and a lot more of the parameters can be determined locally from each other.

The Naïve Bayes model can be incorporated into this: a parent Y gives some probability to its (conditionally independent) children $X_1, \ldots, X_m$. Then, the network structure encodes the assumption

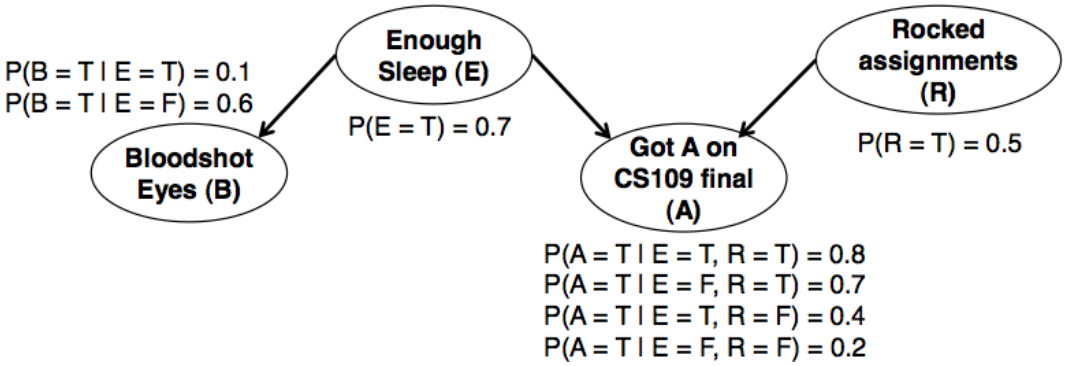
$$P(\mathbf{X} \mid Y) = P(X_1 \cap \cdots \cap X_m \mid Y) = \prod_{i=1}^{m} P(X_i \mid Y).$$

Then, things such as the linearity make more sense. Additionally, if the Naïve Bayesian assumption is violated, then the dependencies between the X_i can be added as arcs, yielding a better model that is still a Bayesian network. But

[38] Does this mean causality or just direct correlation? This is a philosophical problem, and doesn't have a clear answer, given that many physicists claim causality can't really exist.

Bayesian networks are more general than the specific prediction problem: instead of observing all of the X_i, only some subset $E_1, \ldots, E_k$ is observed. Then, the goal is to determine the probability of some set $Y_1, \ldots, Y_c$ of unobserved variables, as opposed to a single Y. Formally, the goal is to find $P(Y_1, \ldots, Y_c \mid E_1, \ldots, E_k)$.

Example 25.1. Consider the following Bayesian network representing conditions around doing well on the CS 109 final:



Here, the professor can know B, A, and R, but doesn't know E. One can calculate $P(A = T \mid B = T, R = T)$ by summing over the unseen variables:[39]

$$P(A = T \mid B = T, R = T) = \frac{P(A = T, B = T, R = T)}{P(B = T, R = T)} = \frac{\sum_{E=T,F} P(A = T, B = T, R = T, E)}{\sum_{E=T,F} \sum_{A=T,F} P(B = T, R = T, E, A)}.$$

Notice how many fewer parameters are necessary for this, and the joint probability decomposes as $P(A, B, E, R) = P(E)P(B \mid E)P(R)P(A \mid E, R)$: the product of the probabilities given the parents.

Another useful concept is a probability tree, which is a tree that models the outcome of a probabilistic event. Probability

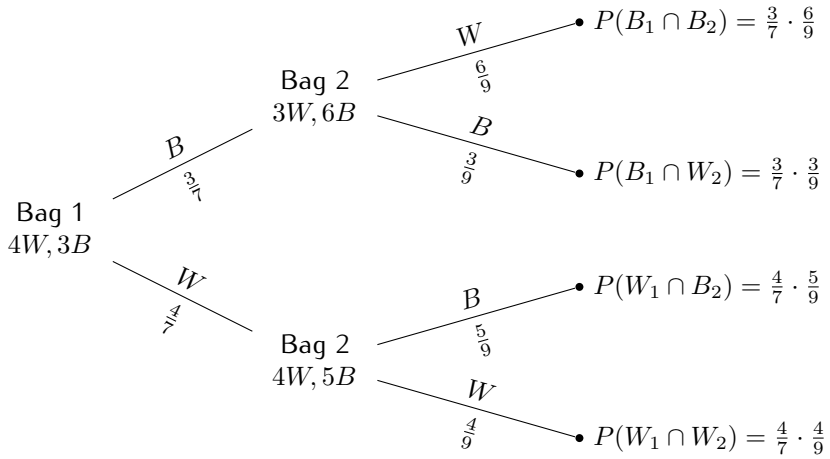


FIGURE 9. An example of a probability tree. Source

trees are traditionally written sideways, and can be thought of as probabilistic flowcharts. This is useful for modelling decisions.

For example, consider a game where one can receive \$10 for not playing, or \$20 with probability 0.5 when playing, and \$0 otherwise. If the numbers are multiplied by 1000, people much more strongly prefer not to play the game. This is an example of utility, as seen before, which means that the leaves of the decision tree are utilities rather than dollar values. This can be useful for making decisions about life expectancy, etc., and are often difficult choices.

In a similar game, the expected monetary value is the expected dollar value of winning, and the risk premium is the amount that one would be willing to give up to eliminate risk: certainty has a utility. This is exactly why insurance

[39] This is the simple algorithm, but not the most efficient one. That's a discussion for CS 228.

companies exist; how much money would you pay to mitigate the existence of a big loss, even if the expected loss is greater.

Utility is nonlinear: suppose one could gain $\$5$ with probability 0.5 and otherwise get nothing, or just receive $\$1$ for not playing. Most people would play, but if the dollar values were scaled by a million, most people wouldn't. In some sense, $\$10000000$ is life-changing, but the next $\$40000000$ isn't. Thus, the shape of the utility curve isn't linear, and depends on one's preference for risk. Most people are risk-averse, so their utility functions are concave down. Some people are risk-preferring, so their utility curves are concave up. A typical utility curve looks something like $U(x) = 1 - e^{-x/R}$. Here, R is one's risk tlerance, and is rough;y the highest value of Y for which one would play a game in which one wins $\$Y$ with probability 0. and loses $\$Y/2$ with probability 0.5, and receives nothing for not playing.

Then, one can calculate how irrational or irrational one is: consider a game in which one receives $\$1000000$ for not playing, and when playing, receives $\$1000000$ with probability 89%, $\$5000000$ with probability 10%, and nothing with the remaining 1%. Similarly, one could choose to play a game where $\$1000000$ is won with probability 0.11 (and nothing happens otherwise), or $\$5000000$ is won with probability 0.1 and nothing is won otherwise.

Sometimes, people pick the second options in both cases, which is inconsistent with any utility function. This is known as the Allais paradox, since the percentages are the same.

A "micromort" is a one-in-a-million chance of death. This leads to an interesting question: how much would one want to be paid to take on the risk of a micromort? How much would you pay to avoid one micromort? People answer differently to these questions. Though this sounds abstract, it comes up in everyday risks that people need to take every day when making decisions. Understanding this makes these decisions a little more insightful.

Finally, consider an example from sports, which has lots of applications of probability. Two batters have their batting averages over two years. If player B has better batting averages each year, he must be better... but what if the first year A has 0.250 and B has 0.253, and in the second year, A has 0.314 and B has 0.321, giving them combined averages of 0.310 and 0.270. Oops; is A better? This is known as Simpson's paradox, and actually happened to Derek Jeter and David Justice in 1995 and 1996. This has applications to marketing (people may prefer different products) or effectiveness of medicine. One might have one medicine favored when the data is broken down by gender, but the other when the total data is considered. This can happen because one treatment has significantly fewer trials than the other in one case, but not the other. Unsurprisingly, this comes up all the time in the real world.

26. Generating Random Numbers: 5/31/13

In many applications, one wants to be able to generate random numbers (and therefore also random permutations, etc.). However, computers are deterministic. Thus, one often settles for pseudo-randomness, which is a condition that a sequence of numbers "looks" random (for a definition that can be made precise; intuitively, it shouldn't be possible to easily predict the next number), but is in fact deterministic.

Most random number generators use a linear congruential generator (LCG), in which one starts with a seed X_0 and gives the next number by a function $X_{n+1} = (aX_n + c) \bmod m$. The effectiveness of this algorithm is quite sensitive to the choices of a, c, and m (e.g. $m = 1$), and often, people take them to be large primes. Generally, language standards don't specify these numbers, or even whether one should use LCG at all, but it's fast and useful.

Notice that the sequence of random numbers must cycle, since there are only m choices for an X_i. This relates to an idea called Kolmogorov complexity.

In C and C++, the `<stdlib.h>` library contains a function `int rand(void)` which returns a value between zero and `RAND_MAX`, inclusive (where the latter quantity is at least 32767 and is nowadays about 2^{64}). Ideally, the values returned are uniformly distributed. Most implementations use an LCG method. The seed can be set using `void srand(unsigned int seed)`, and is often set to the current system time using `seed(time(NULL));`.

Thus, it is a reasonable approximation that `(rand()/((double)RAND_MAX + 1))` $\sim$ Uni$[0, 1)$. Notice the interval is half-open, since zero can be generated, but not 1, and this will be useful in practice.

Then, one has an algorithm for generating a random permutation of a set, such that all permutations are equally likely. Think of this as shuffling a deck of cards.

```
void shuffle(int arr[] int n) {
    for(int i = n - 1; i > 0; i--) {
        double u = uniformRand(0, 1); //u in [0, 1)
        //pick one of the "remaining" i positions
        int pos = (int)((i + 1) * u);
        swap(arr[i], arr[pos]);
    }
}
```

This swaps each element of the array and swaps it with some other random element, which does guarantee tht all permutations are equally likely. Notice that the use of `uniformRand(0,1)` prevents index `n` (which doesn't exist) from being used.

Here's a subtle, but incorrect, variation:

```
void badShuffle(int arr[] int n) {
    for(int i = 0; i < n; i++) {
        double u = uniformRand(0, 1); //u in [0, 1)
        //pick any position
        int pos = (int)(n * u);
        swap(arr[i], arr[pos]);
    }
}
```

Note that this has n^n execution paths, but there are $n!$ permutations. Thus, by divisibility, these permutations can't venly go into the execution paths, so some of them must be more likely than the others.

Here's another choice, which is easier but isn't in-place.

```
void shuffle(int[] arr, int n) {
    double *keys = new double[n];
    for(int i = 0; i < n; i++) {
        keys[i] = uniformRand(0, 1);
    }
    SortUsingKeys(arr, keys, n);
    delete[] keys; //this ain't Java
}
```

Here, `SortUsingKeys()` sorts the array `arr` using the values given in `keys` as weights. Thus, this method is $O(n \log n)$, unlike the first one, which was $O(n)$. Nonetheless, it is still useful in some cases.

Thus, uniform distributions can be made; but what about other ones? A method called an inverse transform offers one solution: if one has a continuous distribution F, let $U \sim \text{Uni}(0,1)$ and define $X = F^{-1}(U)$. Since F is a CDF, then it is necessarily invertible unless it is completely degenerate (in which case we don't really care about it anyways). Then, we have that

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

This last identity is somewhat tricky, but that is the meat of this transform. Thus, X has exactly the distribution we care about. This can also be used in the discrete case, albeit with some modification.

Example 26.1. Suppose $X \sim \text{Exp}(\lambda)$. Then, its CDF is $F(x) = 1 - e^{-\lambda x)}$ for $x \geq 0$. This can be inverted to obtain $x(u) = -\log(1-u)/\lambda$. Thus, $F^{-1}(u) = -\log(1-u)/\lambda$. Since $U \sim \text{Uni}(0,1)$, then $1 - U \sim \text{Uni}(0,1)$ as well, which allows for a simplification $X = F^{-1}(U) = -\log(U)/\lambda$.

This was really nice, but sometimes the closed-form inverse might not exist (particularly in the case of the normal distribution).

Example 26.2. Suppose $X \sim \text{Poi}(3)$. The discrete case is slightly different: given some $U \sim \text{Uni}(0,1)$ and a u generated by that, let x be the smallest value such that $F(x) \geq u$. Since the CDF is increasing, then this is what we want.

There's a simple iterative algorithm to do this, which assumes the existence of a PMF `p(x)` that can be called.

```
int discreteInverseTransform() {
    double u = uniformRand(0, 1); //again, we don't want 1
    int x = 0;
    double F_so_far = p(x);
    while (F_so_far < u) {
        x++;
        F_so_far += p(x);
    }
    return x;
}
```

The last case is that of the normal distribution. Suppose one has a random variable X with a PDF $f(x)$. A method called rejection filtering can be used, which actually simulates another variable with a different density function, though it is at least supported on the same set. Here's the code, and the explanation will follow:

```
double rejectionFilter() {
    while (true) {
        double u = uniformRand(0, 1);
        double y = randomValueFromDistributionOfY();
        if (u <= f(y)/(c * g(y))) return y;
    }
}
```

Here, $c \geq f(y)/g(y)$ for all y so that everything is normalized. Notice that the number of iterations in the loop is $\text{Geo}(1/c)$, which can be unpleasant for large c. Unfortunately, the proof that it works is a bit beyond the scope of the class, though it is in the book.

Example 26.3. Consider simulating the unit normal (since then any other can be created by scaling), which has pdf $f(z) = e^{-z^2/2}/\sqrt{2\pi}$. It can be folded in half, to take the PDF of $|Z|$, obtaining $f(z) = 2e^{-z^2/2}/\sqrt{2\pi}$. This is useful because we already can generate for an exponential distribution using the inverse transform method, so let $Y \sim \text{Exp}(1)$. Thus, $g(y) = e^{-y}$ for any $y \geq 0$. Then,

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2e}{\pi}} e^{-(x^2-2x)/2} = \sqrt{\frac{2e}{\pi}} e^{-(x^2-2x+1)/2+1/2} = \sqrt{\frac{2e}{\pi}} e^{-(x-1)^2/2} \leq \sqrt{\frac{2e}{\pi}} = c \approx 1.32.$$

Thus, $f(x)/cg(x) = e^{-(x-1)^2/2}$, and c is small enough that the loop is mostly efficient. Here, `randomValueFromDistributionOfY()` is just $-\ln(V)$, where $V \sim \text{Uni}[0, 1)$, and the method can be implemented.

Y comes from a half-normal distribution, and can be turned into a normal distribution by flipping a coin to determing the sign of Y to make a normally distributed variable.

Random number generation can be used to numerically approximate integrals, particularly those without a closed form. Intuitively, one takes the graph of the function on the interval and throw darts at it, in some sense, randomy choosing points on the graph. Then, one takes the percentage of the points below the curve and multiply the area in which the points were chosen (the domain of the integral and the maximum value of the function). This is known as Monte Carlo integration, named after a casino town in Monaco.

Example 26.4. As a comparison for accuracy, calculate $\int_0^2 e^x \, dx$. Use the following algorithm:

```
double integrate() {
    //number of "darts" to throw
    int NUM_POINTS = 1000000;
    int numBelowFn = 0;
    for(int i = 0;i < NUM_POINTS; i++) {
        double x = uniformRand(0, 1) * 2.0;
        double y = uniformRand(0, 1) * exp(2.0);
        if (y < exp(x)) numBelowFn++;
    }
    return 1.0 * numBelowFn / NUM_POINTS * 2.0 * exp(2.0);
}
```

Here are some computed values: with 10 darts, the result is about 5.911, and it converges to 6.391 and 6.388 as the number of darts frows to about 10^6. The analytic value is about 6.389, which is pretty nice. Notice that since the Monte Carlo method is random, sometimes it gets worse when `NUM_POINTS` is increased.

Some other methods (e.g. the Runge-Kutta methods) try to distribute the points uniformly, but this requires knowing some things about the function.

Another application of random numbers is to simulate statistical questions.[40] However, one should take care: bugs in the simulation or even coding approximations have compromised some part of the answer.

27. Some Review for the Final: 6/3/13

> "Threee statisticians are out hunting geese. One shoots a goose, but is off by a meter to the right. The second misses by a meter to the left. The third says, 'we got him!' "

The final exam is approaching; here are some topics that will be worth reviewing. Notice that the final covers material from the whole class but is weighted in favor of concepts from the second half.

[40] A good example of this is the problems on our problem sets. Here Dr. Sahami claims that it was a good thing we didn't do this to check our problem sets, though I definitely did...

- Counting and combinatorics.
- Conditional probability and Bayes' theorem.
- Independence and conditional independence.
- Distributions, including the Bernoulli distribution (a coin flip), the binomial distribution (n coin flips), the multinomial distribution (several different outcomes, rather than two), Poisson, geometric (trials until first success), negative binomial (trials until r successes), and hypergeometric (how many balls drawn from an urn are white, without replacement? Note that with replacement, this would just be binomial).
- Distributions, cont.: uniform (equally likely over a range), exponential, and normal.
- Joint distributions and marginal distributions.
- Expectation, conditional expectation, and variance.
- Covariance (not normalized) and correlation (normalized covariance).
- Moment-generating functions. Notice what happens when one multiplies them.
- Inequality-land: Markov's, Chebyshev's, Chernoff's and Jensen's.
- The Central Limit Theorem, which is the big important part of the theory.
- Machine learning: parameter estimation, whether an estimator is biased and/or consistent. Also recall the method of moments and the maximum likelihood estimator (MLE).
- Bayesian estimation, including hyperparameters such as the beta distribution and the Laplace prior (which is a special case of both the beta and the multinomial distributions).
- Relating to machine learning, there are the actual algorithms, such as Naïve Bayes and logistic regression. When should each be used? It's also worth remembering overfitting.
- Finally, the basics of utility theory, which were discussed a couple times in class, might come up.

Here is a much shorter list of things which are *not* on the final exam:

- Bayesian networks are super interesting, but because they weren't written into the homework, they won't be on the final. If you have a burning desire to be tested on these (or to learn more about them because they're cool), then take CS 228 or CS 228T.
- Computer generation of probabilities will also not be tested for similar reasons. Relatedly, Monte Carlo simulation isn't tested (partly because the exam is closed-computer).
- The general Dirichlet and gamma distributions won't be tested, though specific examples of them appear above and will be tested (e.g. the Laplacian).

There is a game called Acey Deucey, in which one takes a standard 52-card deck and flips two cards and then makes the wager that the next card is strictly between the two cards. Notice that an ace is high, but a two is low. What is the probability of winning this wager?

Like many problems, the trick is picking the right variables. Let X be the difference between the ranks of the first two cards. Thus, $P(X = 0) = 3/51$, because after the first card is chosen, there are three more cards that can be chosen. More generally, $P(X = i) = (13 - i)(2/13)(4/51)$. The $13 - i$ term comes because there are multiple ways to get that difference (e.g. 5 and 7, 6 and 8, etc.). Then, there are two ways to draw the first number once the specific pair has been specified by the $13 - i$ term. Then, the remaining card happens four times out of 51.

If Y is the probability that the third card is between the first two, then the goal is to find

$$\sum_{i=1}^{12} P(X = i)P(Y \mid X = i) = \sum_{i=1}^{12} \frac{8(13 - i)}{13 \cdot 51} P(Y \mid X = i).$$

Of the 50 cards that remain, there are four cards of each valid rank, and there are $i - 1$ valid (i.e. winning) ranks. Then, when everything is plugged in, this ends up as about 0.2761.

There's another way to solve the problem, which exploits the symmetry of the problem. Clearly, it's necessary for all three cards to have different ranks, which happens with probability $(48/51)(44/50)$. Then, it is necessary to have the third card in the middle of the first two, which is just a computation on the six permutations of three cards: two of them have the third card in the middle, so the final answer is $(48/51)(44/50)(1/3) \approx 0.2761$. Much simpler.

Another example: in a group of 100 people, let X be the number of days that are the birthdays of exactly three people in the group. What is $E[X]$?

When one sees expected value, the first step is to decompose the problem. Let A_i be the number of people who have a birthay on day i, so that $A_i \sim \text{Bin}(100, 1/365)$. Then, $p = P(A_i = 3) = \binom{100}{3}(1/365)^3(364/365)^{97}$. Thus, let $X_i = 1$ if $A_i = 3$ and $X_i = 0$ otherwise, so that

$$E[X] = E\left[\sum_{i=1}^{365} X_i\right] = \sum_{i=1}^{365} E[X_i] = \sum_{i=1}^{365} P(A_i = 3) = \sum_{i=1}^{365} p = 365p,$$

Notice that the events aren't independent, but in expectation it just doesn't matter.

Another problem: in a group of 100 people, let Y be the number of distinct birthdays. What is $E[Y]$? Define indicator variables for each day: let $Y_i = 1$ if someone has a birthday on day i and 0 otherwise. Then, $E[Y_i] = P(Y_i) = 1 - P(\sim Y_i) = 1 - (364/265)^{100}$, so

$$E[Y] = E\left[\sum_{i=1}^{365} Y_i\right] = \sum_{i=1}^{365} E[Y_i] = 365\left(1 - \left(\frac{364}{365}\right)^{100}\right).$$

But this can be reformulated into a problem about hash tables.

Another question: suppose $X_1, \ldots, X_n \sim \text{Geo}(p)$ are IID random variables. How would one estimate p using the method of moments? $E[X_i] = 1/p$, so $p = 1/E[X_i]$, and one would estimate

$$p = \frac{1}{E[X_i]} \approx \frac{1}{\hat{m}_i} = \frac{1}{\overline{X}} = \frac{n}{\sum_{i=1}^{n} X_i} = \hat{p}.$$

For some retrospective, probability has become incredibly important in computer science. Though computers are deterministic, they are so complex that probability makes a lot of questions easier to answer. This relates hugely to understanding data: the data that users leave on the web provides information that can be used for probabilistic analysis. In some sense, each user is a giant vector with likes, dislikes, written text, etc. Many large companies exist solely to collect and analyze this sort of data. For example, there was a recent controversy in which Target was able to predict whether its users were pregnant (for the purposes of targeted[41] advertising), sometimes better than they were. This is kind of scary, but it all ties back to probability and machine learning. If you want to track how information about you flows, try using different middle intials to see where your junk mail comes from.

This also applies to one's credit card purchases, and loan applications: each purchase (or loan application) is run through an algorithm to determine how likely it is to be legitimate. But in mortgages, all denials are done by a person, in order to eliminate some sort of discrimination. This has connections to artificial intelligence, strangely enough.

One more story: the Department of Defense made a neural network which was to identify tanks in images. It worked perfectly in the lab, but in the field it failed. People had thrown machine learning at the problem without looking at the dataset, because the pictures with tanks were taken on sunny days and those without tanks were taken on cloudy days, and this caused problems. Make sure to understand the problem the machine learning is applied to.

[41] No pun intended.