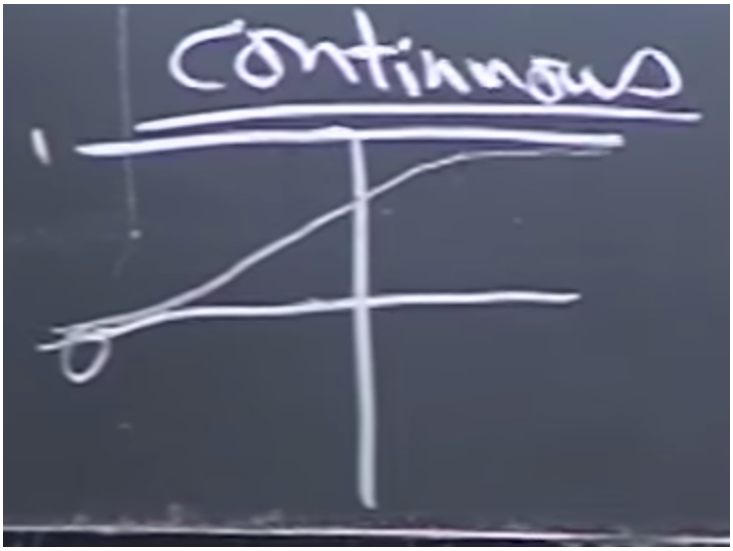
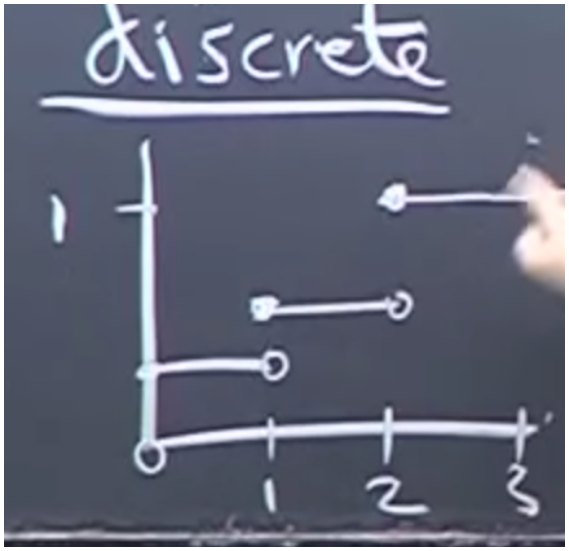
random variables and distributions

- a random variable is a function from a sample space to the real line. So its domain is some sample space and the output of the function is some real number. So basically, it maps some sample space to a real number.
- a random variable is said to have Bernoulli distribution if it can take on 2 possible values (the output of the function is two possible values), 0 and 1. so out of the sample space that's the domain, all outcomes map to either 0 or 1. so if $P(X = 1) = p$ then $P(X = 0) = 1 - p$. note that $X = 1$ and $X = 0$ are events (because these are the resultant mappings of some subset of the sample space, which is what events are: a subset of the sample space).
- binomial, pmf, distribution as blueprint, notation, sum of binomials
- if we have n independent trails of bernoulli and we make a sample space with all possible outcomes, and then make events so that an event is that exactly k of those n trials are successes, that distribution is called a binomial. We can easily show that the binomial distribution is given by $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Notice that $X = k$ is the event (so $P(X = k)$ is the probability of the event (probability that the outcome is one of the outcomes in the subset specified by the event).
- That expression above which gives the probability of an event for a random variable is called the probability mass function. The distribution is just a blueprint that specifies the function defined by the random variable.
- The notation X ~ $Bin(n, p)$ just means that the random variable X takes on the binomial distribution.
- A random variable that has the value 1 or 0, according to whether a specified event occurs or not is called an indicator random variable for that event. We can write the binomial as$X = X_1 + X_2 + ..X_j$ which is the sum of bernoulli indicator random variables.
- Each of the random variables in the sum above are said to be i.i.d → which stands for *independent identically distributed* (it means what it says). A distribution says what is the probability that a random variable will take on this value, that value etc. Two or more random variables can have the same distribution.
- In random variables, when you see $X = 7$ or something like that, its really representing an even (subset of sample space, all of which map to 7). It is not a assignment, equation etc. It is an event.
- Cumulative distribution function (cdf) is the probability of the event $X \leq x$, so cdf is $P(X \leq x)$. cdf is denoted by $F(x) = P(X \leq x)$. This can be defined for continuous and discrete random variables.
- the pmf is a function that specifies the probability for all possible values of the random variable, so $f(x) = P(X = x)$, so $\sum_{allpossiblevalues} pmf = 1$. This is only for discrete random variables.
- To check if something is a valid pmf, you have to check if all values it takes on are ≥ 0 and that they sum up to 1.
- the cdf is more general, the pmf only works in the discrete case.
- either the pmf or the cdf fully specifies a distribution.
- Adding, subtracting, multiplying, squaring, cubing etc are permitted for two functions if their domains are the same. So when we do any of these operation for random variables, it just means adding the output of the functions. Also, note that these operations can only be done if the domains of the random variables (the sample spaces) are the same. This turns out to be important in a later result.
- As we'll see later in the class, adding random variables is called a convolution.
- we will in casual conversation say find the distribution of $X =$ # of something. This is casual talk but notice that the random variable is still a mapping from sample space to a real number. here that real number happens to be the "# of something".

- When you get asked “whats the distribution”, you can find either pmf or cdf, since both are sufficient to describe a distribution. we’ve seen the bernoulli and binomial distributions (we’ve given their pmfs).
- hypergeometric distribution: given a set of marbles with w white marbles and b black marbles, and you pick a simple random sample of n marbles (which means all samples are equally likely), what is the distribution of the number of white marbles k in the sample? This is just $\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$. This is the pmf. This is called the hypergeometric distribution.
- Notice that the difference between the hypergeometric and the binomial distribution is that in the binomial distribution, we kind of pick with replacement, the subsequent probabilities are the same and they are independent, here they are dependent probabilities.
- The cdf in the continuous case is a non-decreasing function (think why). the cdf in the discrete case is a step wise function with breaks, but is still non-decreasing. the cdf will approach 1 as it gets to bigger values and will be 1 at the biggest possible value, if such a thing exists for that distribution.





- Looking at the discrete cdf above, we can recover the pmf by looking at the jump sizes at each integer value. Using this fact, given the cdf, we can compute the pdf function for a given value or a given range of values. In discrete probabilities, watch out for $\leq$ vs $<$.
- A cdf is non-decreasing, right continuous (which means its limit as you approach a point from the right side is equal to the value at that point) and the $lim_{x\to-\infty}F(x) = -1$ and $lim_{x\to\infty}F(x) = 1$. Actually this is an “if and only if” condition for a function to be a cdf.
- Two random variables are independent → what does this mean?. This goes back to the idea that random variables are defined in terms of events. So tying this back to events and independence, random variables are independent if and only if $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ for all x, y. So we’re just back to independence of events. Similar equation for pmf.
- A mean (or average or expected value) for a random variable is defined as $\sum_x x * P(X = x)$. this only works for the discrete case. Note that mean is only one summarizing view of a distribution. It tells us nothing about the spread, tail, weights at different points, etc.
- We can apply the definition and do the math but this is a nice result. Expected value of binomial distribution turns out to be np.
- Expectation of hyper geometric series: ***???*** come back here after doing linearity.
- The geometric distribution Geom(p) is how many failures before the first success when each is a bernoulli with probability p. pmf $P(X = x) = (1 - p)^k p$. When we do $\sum P_k$ over all $k's$ we get 1, as we should for

any pmf. While doing this, we get a geometric series summed to ∞. That's why this distribution is called geometric distribution.

- What is the expected value of the geometric distribution?
  $E(X) = \sum_{k=0}^{\infty} kpq^k = p\sum_{k=0}^{\infty} kq^k$
  Notice that since $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$, differentiating we get $\sum_{k=0}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$. Substituting that back in, we get $E(X) = \frac{q}{p}$.
  Notice that this also lends itself to a recursive expression: Let $f = E(X)$. $f = q + fq$. Solving for f, we get $f = \frac{q}{p}$.
- It is possible to construct random variables where the expectation does not exist because the sum diverges.
- If the expectations exist, linearity of expectation states that $E(X+Y) = E(X) + E(Y)$. This is true irrespective of whether X and Y are dependent or independent! We will prove it for the discrete case here. The proof is analogous for the continuous case. Proof:
  $E(X+Y) = \sum_{s_1}\sum_{s_2}(X(s_1) + Y(s_2))P(s_1 s_2)$, where s_1, s_2 are the elements in the sample spaces that are the domains of X, Y respectively. Now this is
  $\sum_{s_1}\sum_{s_2} X(s_1)P(s_1 s_2) + \sum_{s_1}\sum_{s_2} Y(s_2)P(s_1 s_2) = \sum_{s_1} X(s_1)P(s_1) + \sum_{s_2} Y(s_2)P(s_2) = E(X) + E(Y)$

- Notice how the above proof holds even in the case that X and Y are dependent! This makes it quite a powerful result that we can use to compute the expectations of complicated situations by breaking them down into separate situations each with distribution represented by random variable X_i, such that $\sum X_i = k$ where k the right event that we are interested in.
- There are millions of possible distributions. why are some popular? the ones we are seeing have nice stories behind them that come up in applications often.
- negative bernoulli distribution is the distribution of having r successes. to compute the expectation for this, we should break up the expectation using linearity of expectation. then finding the expectation of each is easy, and summing that up gives us the total expectation.
- remember: this linearity thing is a powerful way to compute expectation. so when asked to compute a complicated expectation, think about breaking it up using the property of linearity.
- A distribution is like a class and each random variable with that distribution is an instance (if you haven't programmed, ignore this sentence)
- The Poisson distribution is given by $P(X = k) = \frac{\lambda^k}{e^{-\lambda}k!}$ for $k = 0, 1, 2, 3...$ and 0 otherwise (so it is only for integers). λ is called the rate parameter and it is any positive real number.
- Lets make sure the given function is valid pmf. We see that it is non-negative. also $e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$, noting that the thing inside the sum is the taylor series for e^{λ}, we get the result 1. So it is a valid pmf. Using the same taylor series observation, we can show that the expected value (mean) is λ.
- the poisson paradigm/poisson approximation is when we have a large number of things that could happen $A_1, A_2, ..A_n$ but each $P(A_j)$ is small (notice it could be different probabilities) but n is large and when each event is independent or "weakly dependent" (we won't define weak dependence mathematically but it roughly means that knowledge of one experiment doesn't affect much the probability of the other), then the number of A_j's that occur follows the poisson distribution approximately. → this is the claim.
- proof: joe doesn't actually prove the derivation for the general case, but does prove how the limit of the binomial is the poisson (see below).
- note that, by linearity of expectation, and treating each of the A's as indicator variables, this means that the mean of the number of things that occur is approximately λ. (since for indicator random variables, their only possible values are 0 and 1).
- the poisson distribution is the single most widely used distribution as a model for discrete data! but why the heck? its used for applications where we want to count the number of occurrences where we have a large number of trials and small probability for each one.

examples:

- number of earthquakes in a given region in a year. why? there are many days in a year each having a very small probability of having an earthquake.
- number of emails you get a day from a friend. you have lots of friends each unlikely to call/email you.
- etc. the pattern is lots of trials, each with low probability.

- notice binomial converges to poisson when n is large and p is small (its a special case where all the p_j's are the same). proof:

  we should choose λ to be the mean, so in the case of the binomial it is np. given that $\lambda = np$, the poisson would predict that the pmf of the binomial in the case of large n and small p is $lim_{n\to\infty, p\to 0}\binom{n}{k}p^k(1-p)^{n-k} = \frac{n(n-1)..(n-k+1)}{k!}\frac{\lambda^k}{n^k}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k}$. Now the right most term is just 1 since $n \to \infty$. The second last term, using limits, is $e^{-\lambda}$. In the first two terms, since $n \to \infty$, the numerator cancels with n^k. This gives us the exactly the poisson pmf formula, that is the limit is the poisson pmf formula.
- a continuous random variable is one which can take on any value in a set of continuous intervals.
- we say that the chance that it can take on any given value is 0, so it doesn't really make sense to talk about the probability mass function for it. we instead talk of something called the probability density function which gives us the measure of instantaneous density at a given point, defined so that $P(a < X < b) = \int_a^b f(x)dx$ where $f(x)$ is the pdf.
- For a pdf to be valid, we must have $\int_{-\infty}^{\infty} f(x)dx = 1$.
- we can estimate the value of of $P(x_0 + \epsilon, x_0, x_0 - \epsilon)$ by $f(x)\epsilon$. (the idea of limits and that the integral is approximated by a sum with small enough ϵ's).
- to get the cdf, we just do $P(X < b) = F(b) = \int_{-\infty}^b f(x)dx$. We call F the cdf, just as in the discrete case. To get back the pdf from the cdf, we just have to differentiate F (fundamental theorem of calculus). Note that either the pdf or the cdf is enough to fully describe a distribution.
- Variance is a measure of spread. We'd like to not use absolutely values since the absolute value function is differentiable at 0 and because the squaring gives us nice properties, so we define it as the expected value of $(X - E(X))^2$, that is variance is $E((X - E(X))^2)$.
- now this messes up the units, so we define something called standard deviation which is the square root of the variance, which brings it back to the original units.
- we see that $E((X - E(X))^2) = E(X^2 + E(X)^2 - 2XE(X)) = E(X^2) - E(X)^2$.
- The most basic continuous distribution is the uniform distribution which has the same pdf throughout the interval. For this to be a valid pdf, we will need $\int_a^b cdx = 1$, which means $c = \frac{1}{b-a}$.
- expected value of a continuous distribution is defined as $E(X) = \int_{-\infty}^{\infty} x * f(x)dx$.
- the expected value and variance of the uniform distribution turn out to be $\frac{a+b}{2}$, $\frac{1}{12}$ respectively (this is easily shown).
- law of the unconscious statistician (lotus) helps us find the expected value of a function of a random variable (note that the function of a random variable is itself a random variable, but not necessarily with the same distribution, unless it is 1:1 function). $E(g(X)) = \int_{-\infty}^{\infty} g(X)f(x)dx$ (you would think this is too good to be true but it actually is true), think about why.
- we can use the random variable to simulate any distribution; this is called the universality of the uniform distribution and is applicable to software simulations since computers can give us a uniform random number (uniform distribution) and given this and a cdf, we can implement a function that gives us a random variable with the given cdf. let $F(x)$ be the cdf we want to simulate. for now, we will assume F is increasing and continuous (just to make the proof easier, it can also be shown in the general case). Because F is increasing, we can define the inverse and let $X = F^{-1}(U)$. And now, X ~ F (X is distributed according to F) → think about why this is true.

- linear transformations of the uniform distribution gives us uniform distributions. eg. $a + bU$ where U is a random variable with uniform distribution.
- independence of random variables is defined with the same notion as independence of events. just remember that each $P(X \leq x)$ just describes an event. so with this in mind, random variables are independent if and only if:

  $P(X_1 < x_1, X_2, x_2..X_n < x_n) = P(X_1 < x_1)...P(X_n < x_n)$ for all $x_1...x_n$

  the lhs is called the joint distribution and each terms in the rhs the marginal distribution, as we'll see later.

  Note that this also means that it is pairwise true, any combination of true.. etc.
- note that independence of a set of random variables implies pairwise independence but pairwise independence does *not* imply independence.
- the normal distribution (also called gaussian, but he didn't use it first so we won't call it that) is probably the most important distribution - one of the reasons its important is because of the central limit theorem result which we will see later.
- the basic form of the normal distribution has a pdf $\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$. Integrating this, by change of variables from $-\infty$ to ∞, we get 1, which shows its a valid pdf (the $\frac{1}{\sqrt{2\pi}}$ is to make the pdf integrate to 1). look at the graphs of $e^{\frac{-x^2}{2}}$, $e^{\frac{-x^2}{4}}$ etc to get a feel of the shape and how it changes. (the squaring of the x makes the graph symmetric about the y axis).
- By symmetry of the graph, the mean (expected value) is 0 using the property that $\int_{-a}^{a} f(x)dx = 0$ whenever $f(x)$ is an odd function. (the integral that's 0 is $\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx$)
- Variance is $E(X^2) - E(X)^2$ but $E(X) = 0$, variance is just $E(X^2) = \int_{-\infty}^{\infty} x^2\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx$. By integration by parts, we can show that the variance turns out to be 1.
- ϕ is the notation for the standard normal distribution (with mean 0 and variance 1). so $\phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$. Since these is so popular tables exist to compute this easily.
- By symmetry, $\phi(-z) = 1 - \phi(z)$.
- As we'll see later, $E(X)$ is called first moment, $E(X^2)$ is called second moment, $E(X^3)$ is third moment, etc. For the normal distribution, odd moments are 0 since odd moments turn out to be odd functions.
- Lets say we have a random variable, *any* random variable X with mean $E(X)$. If we take another random variable $X + a$ where a is a constant, note that the mean is just $a + E(X)$. The variance, however, is unchanged by adding a constant.
- Now if we multiply by a constant k, the mean is $kE(X)$ and the variance is $k^2V(X)$. mean gets multiplied by k but variance by k^2. To see why, remember that variance is $E(X^2) - E(X)^2$.
- so if we take the standard normal distribution with mean 0, variance 1, and we make up a new random variable $X = \mu + \sigma N$ where N is the standard normal, we see that the mean gets shifted by μ and the variance becomes σ^2. (standard deviation is σ).
- note that the sample space domain of X and of N are exactly the same. The mean and the variance change because the definitions of the mean and variance make them depend on the value of the random variable (the range/output of the random variable function). in other words, think of X as just a linear function of the N. the random variable maps from sample space to N and the linear transformation maps N to X by doing $\mu + \sigma N$. Given this, to get the cdf of X, just replace x by $\frac{x-\mu}{\sigma}$ in the cdf expression for N. We've seen that differentiating the cdf gives the pdf, so on differentiating the cdf, we get the pdf to be $\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.
- The 68-95-99.7% rule is the fact that if X ~ $N(\mu, \sigma)$, then the chance that x is within 1, 2 and 3 standard deviations from the mean μ is 68%, 95%, 99.7% respectively.

- Back in the poisson, we didn't find the variance of it. To do it, note that we need to find $E(X^2)$ which is $e^{-\lambda}\sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!}$. Using the taylor expansion formula for the pmf and differentiating twice on both sides, we see that the variance turns out to be λ. So poisson has mean λ and variance λ!
- Similarly, for binomial, we can find $E(X^2)$ using the fact that it is broken into a sum of indicator random variables, each of which is independent and using symmetry, we can show that the variance is $np(1-p) = npq$.
- to prove lotus (we'll only do it for the discrete case here but the idea is the same for the continuous case), we need to think of the fact that to compute an expectation, we can either sum over all values of the random variable or we can sum over the sample space. this is really why lotus is true. that is to compute $E(g(X))$, the definition is $\sum_{g(X)} g(X)P(X = g(X))$. but this is the same as summing over the sample space, so $\sum_x g(X)P(X = x)$ → think about why this is true.
- the exponential distribution has pdf $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, and 0 otherwise. We can see that the integral of this from $0 \to \infty$ is 1. We see that the cdf is $F(x) = 1 - e^{-\lambda x}$.
- The standard exponential is when $\lambda = 1$.
- why is the exponential distribution important? one reason is it has the memoryless property. this means that if some event follows a distribution with this property, then $P(X > s + t | X > s) = P(X > t)$. We can show pretty easily, from the definition of the conditional probability (bayes' theorem) that the exponential distribution is memoryless. But think about how interesting this result is. The probability that you have to wait at least an additional t after having waited s is the same as the probability of waiting a t from the beginning. so waiting for the s seconds didn't change anything. hence the name memoryless
- the discrete analog of the memoryless is the geometric distribution. so geometric and exponential are closely related.
- conditional expectation of a random variable X is something like $E(X|X > a) = \sum_x x * P(X = x | x > a)$.
- Using the fact that a memoryless distribution is memoryless (obviously), we can show that the only distribution that is memoryless is one in the exponential distribution. To prove this, let $F(x)$ be the cdf of some memoryless distribution. Now $P(X > x) = 1 - F(x) = G(x)$. Memoryless says that $G(x + t) = G(x)G(t)$ (think about why). From this, we can show that solve for G and show that it has to be of the form $e^{-\lambda x}$, which means $F(x) = 1 - e^{-\lambda x}$.
  this shows us that exponential is the only memoryless distribution.