more about distributions

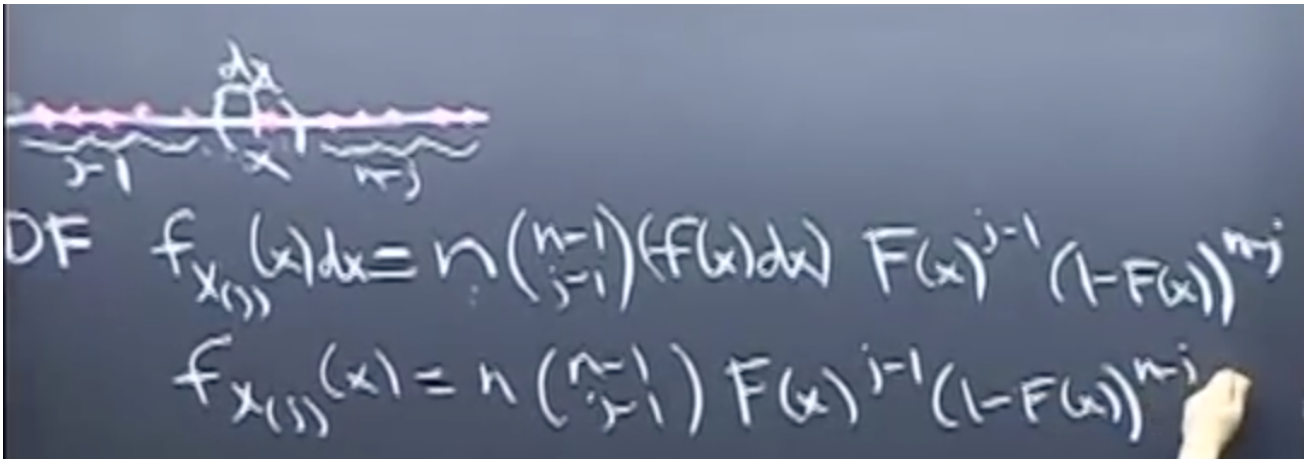
- Just like pdf, cdf etc, a moment generating function (mgf) is a function used the describe a distribution.
- A random variable has moment generating function $M(t) = E(e^{tX})$ (a function of t), if this mgf is finite some interval $t \epsilon (-a, a)$.
- Using the taylor series for e^x, we get $M(t) = E(e^{tX}) = E(\sum_{n=0}^{\infty} \frac{x^n t^n}{n!})$. Now, where ever the sum is finite, we can interchange E and the $\sum$ to get $\sum_{n=0}^{\infty} \frac{E(x^n)t^n}{n!}$. This has the $E(x^k)$ terms which are called the moments. (and hence the name moment generating function). $E(X^k)$ is called the k^{th} moment.
- why are MGFs useful?
  - MGF gives us the moments. That is the the coefficient of $\frac{t^n}{n!}$ in the taylor series expansion gives us the moments of X. These moments also happen to be the derivative of M evaluated at $x = 0$, like so $M^{(n)}(0) = E(X^n)$ → that is, to get the n^{th} moment, we take the n^{th} derivative of the MGF at 0.
  - the MGF uniquely identifies the distribution (if two random variables have the same MGF, then they have the same distribution). → "this fact is very difficult to prove so i'm not going to try to prove it" - Joe Blitzstein.
  - MGFs help with convolutions
- From a computational standpoint, we've seen MGFs. But Joe doesn't do a great job of explaining why we need MGFs. He mentions they can be used in convolutions of independent distributions but doesn't prove it. He uses them to find moments but doesn't say why we need moments or why moments are useful. But we've seen the definition of MGF.
- Joe derives the MGFs for the bernoulli, binomial and the normal distributions. nothing unusual here, but there was a completing the square trick required for the normal distribution's mgf. also, in the derivation, we get a general expression in terms of t, so we can use this result to compute the mgf for any t.
- the posterior distribution means the distribution after we collect the data. the prior is what is known before. so the posterior is something like $P(x|a)$ whereas the prior is $P(x)$. The relationship between these, using bayes' rule is:
  $P(x|a) = \frac{P(a|x)P(x)}{P(a)}$ Note that often, we wont' know the denominator but we can find out the denominator $P(a)$ using the law of total probability → $P(a) = \sum_x P(a|X = x)P(X = x)$.
- Given at least two random variables *X*, *Y*, ..., the joint probability distribution for *X*, *Y*, ... is a distribution that gives the probability that each of *X*, *Y*, ... falls in any particular range or discrete set of values specified for that variable.
- The joint cdf/pmf in the two random variable case is: $F(x, y) = P(X < x, Y < y)/$ $F(x, y) = P(X = x, Y = y)$. In the case X, Y are independent, these probabilities just boil down to products → $P(X < x)P(Y < y)/P(X = x)P(Y = y)$. Note that for the continuous case, the cdfs are multiplied, and *not* the pdfs.
- the marginal distribution just means the random variable taken alone, so like $F(x) = P(X < x)$. In words we can say that independence means that the joint cdf/pmf is just product of the marginals.
- In the two random variable case, the joint pdf $f(x, y)$ is such that to find the probability that (X, Y) lie in a certain region (note that its not a part of a line segment anymore, its a 2-d region), we just do $\int \int_R f(x, y)dxdy$ where R is the region of interest.

- Given the joint distribution of X, Y if we want to find the marginal of X we just do $P(X = x) = \sum_y P(X = x, Y = y)$. Similarly in the discrete case, $f(x) = \int_{-\infty}^{\infty} f(x, y)dy$. (this is the pdfs)
- just like we took the derivative to get from cdf to pdf, in the joint case, to go from cdf to pdf we do $f(x, y) = \frac{\partial}{\partial x \partial y} F(x, y)$ and it follows that to go back to cdf we do$F(x, y) = \int \int_R f(x, y)dxdy$.
- To do conditionals in the joint case, we do $f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$. Note that these are all actually defined for cdfs, but we can show that the above are true (in terms of pdfs) by taking partials as shown before.
- lotus holds in the 2-d case too. $E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$ → its completely analogous, for easily provable reasons.
- if X, Y are independent, then $E(XY) = E(X)E(Y)$.
  proof:
  we are just going to use 2-d lotus here and use the fact that independence means $f(x, y) = f(x)f(y)$ and define the function g as $g(x, y) = xy$. so we get $E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x)f(y)dxdy = \int_{-\infty}^{\infty} xf(x)dx \int_{-\infty}^{\infty} yf(y)dy = E(X)E(Y)$hence proved.
- we'll see later that uncorrelated means that $E(XY) = E(X)E(Y)$. so now we've seen that independence implies uncorrelated.
- MGF of a sum of independent random variables is equal to the product of their MGFs. (see why → this is pretty easy to prove)
- Using the above nice property of MGFs, we can show that the sum of two normals is a normal. with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.
- the multinomial distribution is similar to binomial. represented by Mult $(n, \vec{p})$ where $\vec{p}$ is a probability vector (contains non-negative numbers that add up to 1 → each represents a disjoint case). n is the number of things that are being classified into one of k categories. we have random variables $X_1 ... X_k$, where X_i is the number of objects classified into the i^{th} category. The multinomial is the probability that $X_1 = n_1, ..X_k = n_k$ where $n_1 + n_2 + ...n_k = n$. the probability of any particular assignment of objects to categories is ($n_1, n_2..n_k$ is $p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$. but there are many permutations that can give $X_1 = n_1, ..X_k = n_k$, so considering all the permutations, the answer is $\frac{n!}{n_1! n_2! ... n_k!} p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$
- the above is a joint distribution. now we want to find the marginal. that is, we want to find $P(X_i = x)$. now this is just binomial! → think about why.
- covariance of two random variables is $E((X - E(X))(Y - E(Y)))$. the intuition is that if when X is above its mean, then Y is also above its mean, then this number goes high and same if both are below their means. if its the opposite, then the product is negative and the number becomes lesser.
- note that when $X = Y$, covariance is just variance since variance is $E((X - E(X))^2)$. so cov (X, X) = var (X)
- note that it is symmetric so cov (X, Y) = cov (Y, X) and that cov (X, c) = 0 (c is a constant).
- note that $E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$ → super easy to prove, just multiply out and use linearity of expectation.
- Cov (cX, Y) = c*Cov(X, Y)
- also cov (X, Y + Z) = cov (X, Y) + cov (X, Z)
- cov (X + Y, W + Z) = cov (X, W) + cov (X, Z) + cov(Y, W) + cov (Y, Z)
- by applying the above property, we can show more generally that $Cov(\sum a_i X_i, \sum b_i Y_i) = \sum_{ij} a_i b_j Cov(X_i Y_j)$
- We can show that for independent variables, $var(X + Y) = var(X) + var(Y)$. proof: $var(X + Y) = cov(X + Y, X + Y) = cov(X, X) + cov(Y, Y) + 2cov(X, Y) = var(X) + var(Y)$

(since cov (X, Y) = 0 for independent random variables.)

- also note that var (X - Y) = var (X) + var(Y).
- note that unless random variables are independent, we can't for sure say that $var(X+Y) = var(X) + var(Y)$.
- uncorrelated means that the covariance is 0. independent → covariance 0 → uncorrelated.
- It is *not true* that if the covariance is 0, they are independent. there are many dependent random variables that end up having 0 covariance.
- correlation of X, Y = $\frac{cov(X,Y)}{sd(X)sd(Y)} = cov(\frac{X-E(X)}{sd(X)}, \frac{Y-E(Y)}{sd(Y)})$.
- note that written in the last form, each random variable we are taking the covariance of has mean 0 and variance 1! because of this, correlation is always between [-1, 1]. btw, this thing we did here to make it mean 0 variance 1 is called standardization.
- proof that it is in [-1, 1]: hint: expand out var (X + Y) and var (X - Y) where X and Y are standardized and note that both are greater than 0.
- WLOG → without loss of generality
- so correlation (X, Y) is just standardizing X and Y and then taking the covariance of the standardized things.
- a transformation is just a function of a random variable. for transformations that are differentiable everywhere and strictly increasing, we can write down the cdf as $P(g(X) < y) = P(X < g^{-1}(y))$. differentiating to get the pdf, we get $f_Y(y) = f_X(x)\frac{dx}{dy}$ (using the chain rule this is pretty easy to show).
- Joe mentions the version of transformations for random variable vectors that uses jacobians (the determinant of the matrix that contains all combinations of partial derivatives of one vector with respect to another) but doesn't prove it. no proof means no notes.
- convolution of random variables means the distribution of a sum of indep. random variables. in the discrete case, one way to find this is by conditioning. Let $T = X + Y$.
  $P(T = t) = \sum_x P(X = x)P(Y = t - x)$.
- Similarly, in the continuous case, we can find the pdf of $T = X + Y$ by doing something like $F_T(t) = P(T < t) = \int_{-\infty}^{\infty} P(X + Y < t)f_x(x)dx = \int_{-\infty}^{\infty} F_Y(t - x)f_x(x)dx$. differentiating this with respect to t we get the pdf.
- notice that the only continuous distribution we've so far that is bounded is the uniform. normal and exponential went to ∞. now we introduce another one: the beta distribution. it is given by pdf $f(x) = cx^{a-1}(1-x)^{b-1}, 0 < x < 1$ and $a > 0, b > 0$.
- notice that beta is more a family of distributions/template for a distribution. as we vary a, b we can get different distributions. eg. $a = 1, b = 1$ gives us the uniform distribution. we can give different values for a, b and get different pdfs.
- we can show that the beta distribution is a "conjugate prior" to the binomial. what does this mean? it means that if we compute the distribution of binomial, but with unknown p (in Bin(n, p) and start with the beta as the prior distribution for the p, and then find the posterior $P(p|X)$, we can show using bayes' rule that is also beta. hence the name conjugate prior.
- how would we do this integral in a story way? $\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx$? bayes' billiards: say you have n +1 balls and you paint one of them (at random) pink and drop the balls uniformly randomly between 0 and 1 on the number line. what is the chance that the number of white balls before the pink ball is k? since this is equivalent to spreading the n + 1 balls between [0,1] with uniform probabiliy and then picking a ball to paint, the probability of k balls being before the pink ball is $\frac{1}{n+1}$. this is the same as the sum of probability of (rolling the pink ball first and having it stop at x and with x the binomial probability, looking at how many white balls stop before x). So that integral there is just $\frac{1}{n+1}$.

- Notice that the above is a special case of the beta distribution where $a-1=k, b-1=n-k$.
- A financial derivative is a function of a random variable, often a financial asset. For example if S represents the price of Google stock (the random variable or the financial asset), then $g(S)$ is a financial derivative. g could be something like "pay me if the input is larger than $500 exactly 3 months from now. so a financial derivative is also a random variable (since it is a function of a random variable). you might compute expectations to see if it is good derivative (contract to get into) etc.
- the gamma function is $\Gamma(a)=\int_0^\infty x^a e^{-x}\frac{dx}{x}$. Doing integration by parts, we see that $\Gamma(x)=x\Gamma(x-1)$. notice the similarity between this and the factorial function. but this is a continuous function. so this lets us compute "factorial of fractional numbers" like $\pi!$.
- The Gamma distribution is by pdf $f(x)=\frac{1}{\Gamma(a)}x^a e^{-x}\frac{dx}{x}$. obviously, integrating this gives 1, because we normalized with $\Gamma(a)$.
- you can prove that if the number of emails in a given time is poisson, then the time to first email is exponential and the time to second after receiving the first is also exponential (by memoryless property), but what is the time to n^{th} email? this is same as a convolution of the time to i^{th} email after the $i-1^{th}$ has arrived. we can show that the convolution of the exponentials is the gamma.
- here, we will do a proof using MGFs:
  the mgf of $T(i)-T(i-1)=\frac{1}{1-t}$ (because this is just exponential). then by convolution, $T(n)=(\frac{1}{1-t})^n$ But we don't know yet that this is the mgf of the gamma distribution. using lotus to find the mgf of the gamma distribution, we see that the mgf of gamma is indeed $T(n)=(\frac{1}{1-t})^n$
- given we have the MGF, we can find any moment. technically, here we did rely on the fact that MGF same → distribution is the same, which Joe didn't prove but stated.
- Order statistics: Let $X_1, ...X_n$ be i.i.d. The order statistics are $X_{(1)}\leq X_{(2)}...\leq X_{(n)}$ where $X_{(1)}=min(X_1,..X_n)$ and $X_{(n)}=max(X_1,..X_n)$. (here $X_{(j)}$ is a number generated by the X_j) the median is the middle number when there are a odd number of total elements in sorted order. Similarly, we have 25th percentile, 75th percentile, interquartile range which is difference between 75th and 25th percentile. we'll find the distribution for these things, since they are not fixed numbers from a sample set.
- "the discrete case is tricky and we will only talk about the continuous case" - hopefully we'll see why soon.
- finding $P(X_{(j)}\leq x)=P(\text{at least } j \text{ of the } X_{(i)}\text{'s} \leq x)=\sum_{k=j}^n \binom{n}{k}F(x)^k(1-F(x))^{n-k}$, and on differentiating we get the pdf for the j^{th} order statistic.
- the other way to find this pdf:



way to find the pdf of the order statistic without first finding the cdf and then differentiating.

the above uses the idea of infinitesimal probability within a region dx.

- similarly, we can find the joint pdf of two order statistic. we can draw a similar function but with two such dx regions.
- Notice that the pdf above is a beta!