more topics in probability and stats

- Conditional expectation in the discrete and continuous case:
  $E(Y|X=x) = \sum_y yP(Y=y|X=x)$ in the discrete case and $E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)}dy$ in the continuous case.
- $E(Y|X=k)$ is basically saying that $E(Y|\ event\ X=k)$. $E(Y|X)$ will be a function of x. If we plug in a given k to that function, we get $E(Y|X=k)$.
- some properties of conditional expectation:
  - $E(h(X)Y|X) = h(X)E(Y|X) \to h(X)$ is a constant since here we are basically treating X as a given ($X=k$)
  - $E(Y|X) = E(Y)$ if X, Y are independent. Note that the converse isn't true.
  - Iterated expectation (or Adam's law) says that $E(E(Y|X)) = E(Y)$. This is basically the same as the law of total probability. It is frequently used to compute $E(Y)$ as $E(E(Y|X))$ (proof idea: conditioning, iterating, those kind of ideas)
  - $E((Y - E(Y|X))h(X)) = 0$ proof:
    $E((Y - E(Y|X))h(X)) = E(Yh(X)) - E(E(Y|X)h(X)) = E(Yh(X)) - E(E(Yh(X)|X)) = E(Yh(X)) - E(Yh(X)) = 0$
    (the last step uses iterated expectation).
- conditional variance is defined similar to variance, just that we treat everything as thought we know X. So the variance is $E((Y - E(Y|X))^2|X) = E(Y^2|X) - E(Y|X)^2$ (this equality is just like we proved in the "non-conditioned" version.
- Now $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$. proof is left as an exercise. its pretty easy.
- inequalities are good because we can get exact bounds and ranges on things. approximations are useful but inequalities let us really understand how close we are.
- some inequalities:
  - cauchy schwartz: we have already proven this when mean is 0 (go back to correlations). joe does not prove it for the general case 😓
  - jensen's inequality:
    if g is convex ($g'' > 0$), $E(g(X)) \geq g(E(X))$. the inequality flips for concave functions.
    proof: stare at this for a while:

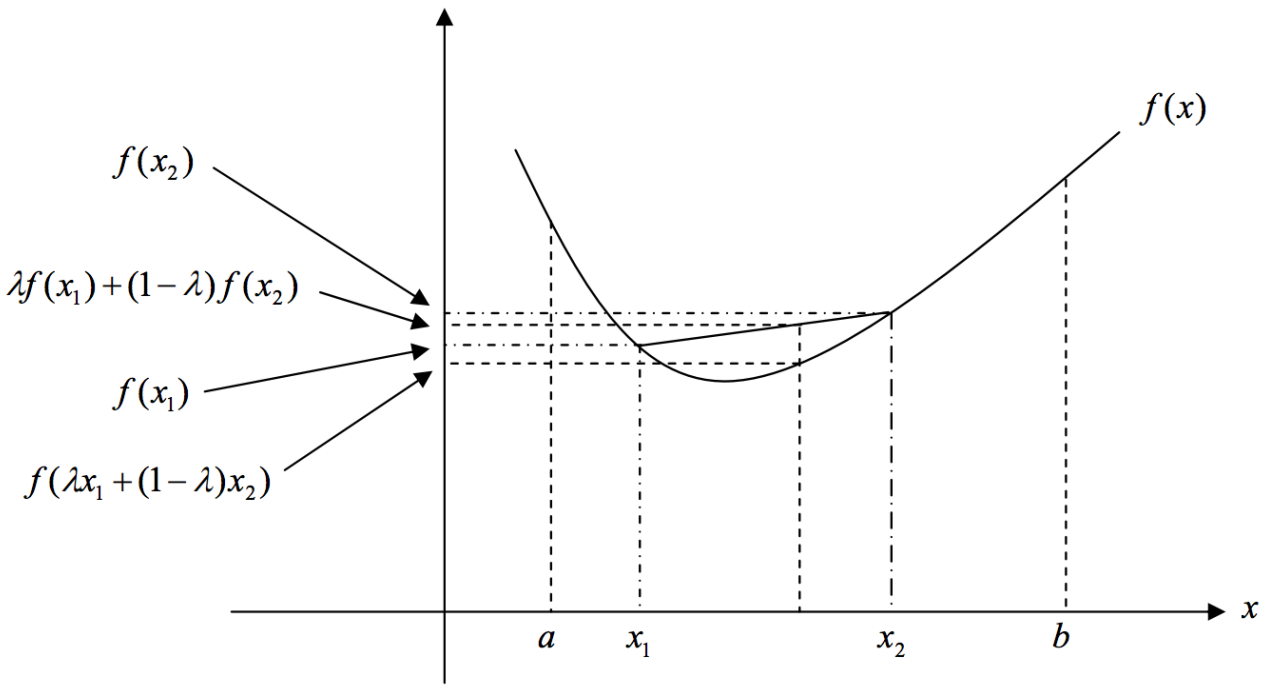
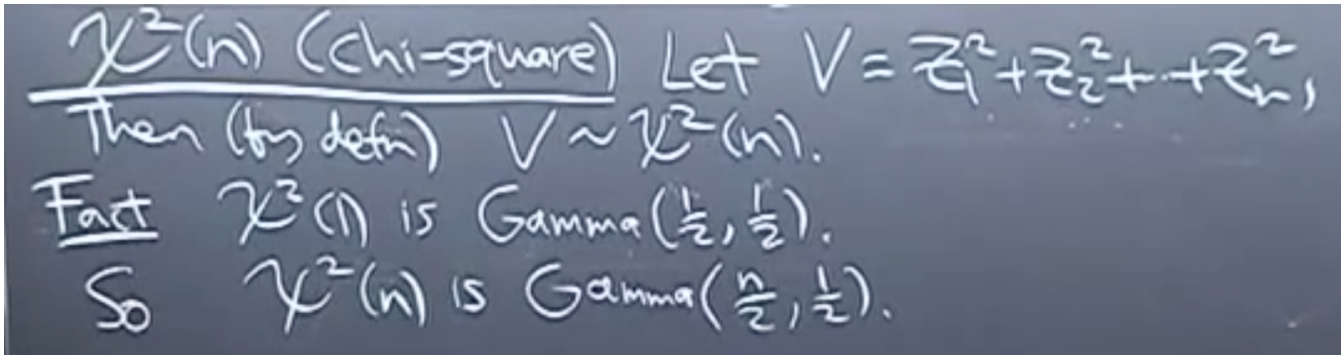


Figure 1: Illustrative example of convexity.

- Markov's inequality:
  $P(|X| \geq a) \leq \frac{E(|X|)}{a}$ for all $a > 0$.
  proof: $P(|X| \geq a) \leq \frac{E(|X|)}{a}$ is the same as $aP(|X| \geq a) \leq E(|X|)$ → to see why this form is true, just write the formula for $E(|X|)$ and you will see why.
- Chebyshev's inequality:
  $P(|X - \mu| > a) \leq \frac{Var(X)}{a^2}$ for all $a > 0$
  proof: write the above as $a^2 P(|X - \mu| > a) \leq Var(X)$ and think about the definition of $Var(X)$.

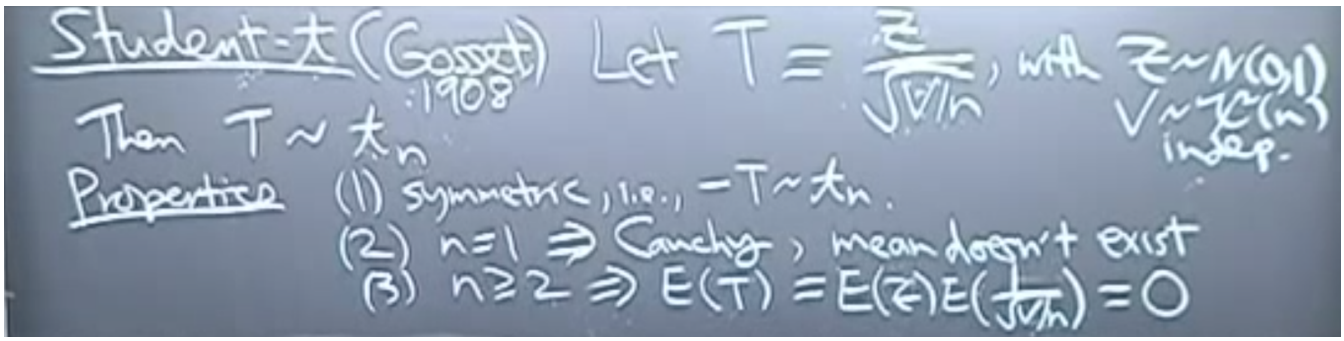
- The Chi-square distribution:
  Note that z_i's are i.i.d standard normal.



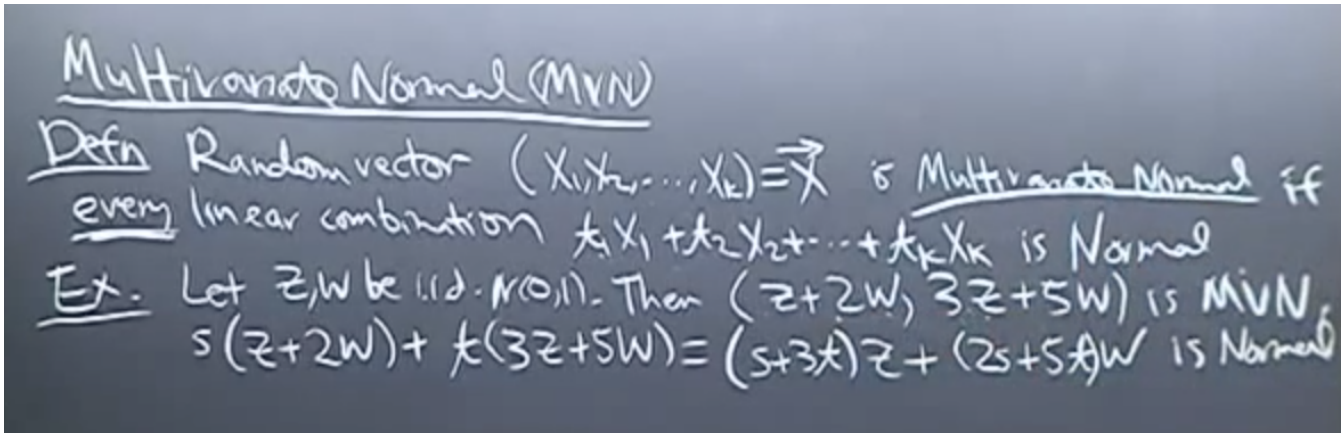
It can be shown that V is Gamma (1/2, 1/2).

The Chi square shows up in statistical tests wherever we add up squares of random variables that are normally distributed (which turns out to be a lot of places).

- we have derived things for the gamma which all apply here. moving on.
- The student-t distribution:



- Multivariate normal:
  Recall that we showed using MGFs that adding independent normals gives us a normal distribution. Having said that, here's the definition:



- the pdf of multivariate normals uses jacobians and stuff and Joe Blitzstein didn't do it 😕
- Note that to find the MGF of the multivariate normal, all we have to do is use the fact that $t_1X_1 + .. + t_nX_n$ is a univariate normal and then just apply the MGF result for univariate normal.
- Law of large numbers:

  Let $X_1, ...X_n$ be i.i.d with mean μ and variance σ^2. Let $\overline{X}_n = \frac{\overline{X}_1 + ... + \overline{X}_n}{n}$. The law of large numbers says that $\overline{X}_n - \mu \to 0$ as $n \to \infty$.
- Intuitively, if we go back to the definition of a limit, what the law of large numbers is really saying is that given a small enough target error between the observed mean and the theoretical mean, we can give a n such that probability of the observed error being within c is as arbitrarily close to 1 as we'd like.
- Unfortunately, Joe does not give a satisfactory proof for the law of large numbers. but we can see that this is super important because if law of large numbers weren't true, what basis/hope would we have to repeat experiments a large number of times and come to the conclusion that we have a reasonable measure of some quantity. This is saying "you get more and more data, you will converge to the truth. Without this law, many scientific experiments would be hopeless."
- the above does not tell us anything about the distribution of the sample mean (or observed mean), $\overline{X}_n$. The central limit theorem has something to say about that.
- $\sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \to N(0, 1)$ as $n \to \infty$. Note that $\sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma}$ is the sample mean. so really, the central limit theorem gives us the distribution of the sample mean as $n \to \infty$, which is standard normal.
- There are two things to note:
  - this is saying that the sample mean converges to standard normal even if the original distribution its drawn from is not normal or even related.
  - note that it is saying that the distribution converges. this is different from the probability converging, like we saw in the law of large numbers.
- proof idea of central limit theorem:
  - using the idea that same MGF implies same distribution (which Joe didn't prove in the lectures), we compute the MGF of the above quantity and it ends up being the same as the MGF of the standard normal. Hence the conclusion.
- One of the applications of this theorem is to approximate any distribution by a normal, eg a binomial by a normal, with the assumption that n (the number of trials) is large. This assumption is needed to bound the error properly.
- adding "gaussian" noise makes a lot of sense because of the central limit theorem which says things about samples, averages and and the distribution of how the sample is off the average.
- it might seem weird to approximate discrete distributions with continuous distributions because $P(X = a)$ is 0 for a continuous distribution. As an approximation, we do something called a continuous correction where for $P(X = a)$, we approximate it by $P(a - \frac{1}{2} < X < a + \frac{1}{2})$.