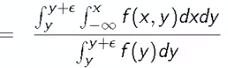
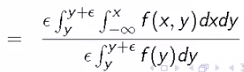
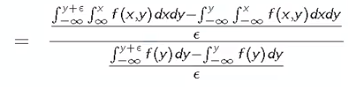
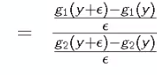
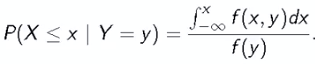
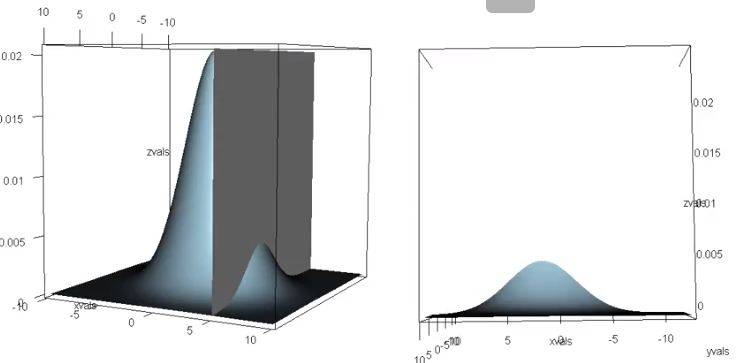
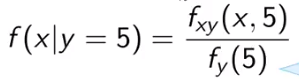
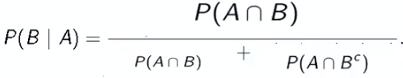
**Conditional Probabilities and Densities**

* P(1) when rolling die = assumed to be 1/6, but suppose we had extra info that the roll ended in an odd
* **Conditional on this new info,** p(1) now = 1/3
* Let B = event w/ P(B) > 0
* Conditional Prob of A given B has occurrent =  (intersection divided by given)
* **If A and B are *independent*, then** (P(B)’s cancel)
* For the die roll, B = {1,2,3}, and A = {1}, so P(A|B) =  🡺 ***A is a subset of B, so the intersection is P(A) by itself*** 🡺 P(A)/P(B) = 1/6 / 1/2 = 2(1/6) = 1/3
* **Conditional Densities/Conditional Mass Functions =** densities + mass functions that govern the behavior of a random variable conditional on the value another random variable took (i.e. **functions of 1 variable conditional on the value of another**)
* f(x,y) = **bivariate (joint) density** or **mass function** for random variables X and Y 🡪 f(x,y) governs the probabilistic behavior of the random variables
* Then let f(x) and f(y) be the associated **marginal mass function** or **densities** that are *disregarding the other variables*
* Then, 
* continuous = integral of joint density f(x,y) over x
* discrete = joint mass function f(x,y) summed over x
* **In other words, to know the marginal probability behavior of random variable y (regardless of what happened w/ respect to x), integrate over all possible values of random variable x**
* Do the flip-side for marginal probability behavior of random variable x
* Then, **conditional density** or **mass function** *given Y = y* is given by  (joint / marginal, from the previous definition)
* Easy to see that (for discrete/finite cases), definition of conditional *probability* is the same as the one for conditional *events* where A == event that X = x and B == event that Y = y
* Continuous definition = bit harder, since events X = x and Y = y have probability = 0
* Useful motivation = take appropriate limits as follow:
* Define:  🡪 random variable X is less than or equal to some real value x, and random variable Y lies in some interval from real value y to y plus some error
* Now both A and B = events w/ probability > 0, and we can apply the standard definition of conditional probability to get P(A|B)
* 
*  🡪 sub in our continous definitions for A and B
*    where 
* We notice that the limit of the numerator and denominator both trend towards g1’ and g2’ as epsilon gets smaller and smaller (*closer to conditioning Y being some specific value y*)
* This gives the **conditional distribution function** associated w/ x as 
* Taking the derivative w/ respect to x gives the **conditional density =** 
* **Densities = derivatives of distribution functions**
* Geometrically, conditional density is obtained by taking the relevant slice of the joint density f(x,y), which itself is a surface (*with volume w/in = 1 to be a valid joint density*) where (x,y) = the plane, f(x,y) = z = the height
* At a particular value for Y, we get some plane (say 5), which slices throught he surface and gives some function

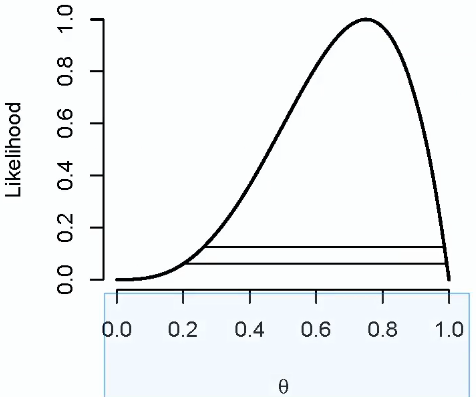
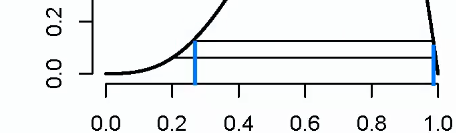
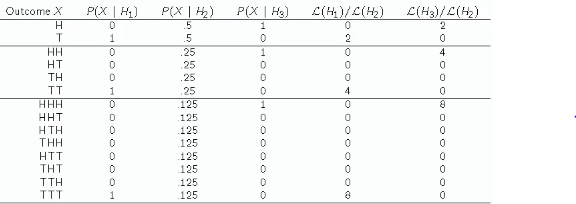
 🡺 

* This slice/function on the right, does not integrate to 1, so it’s not a valid density
* So, we take the slice + appropriately **renormalize** it by something that integrates to 1, say 
* This gives 
* This idea extends to any other line, even non-linear functions)
* Ex: 
* NOTE: **marginal density of y =** integrate joint density function over x 🡺 
* Therefore: **conditional density** ==  == conditional density function/governing behavior of random variable X given y = 3, we have **3\*e^(-3x)**

**Bayes' Rule and DLR’s**

* f(x|y) = conditional density/mass function for X given Y = y, then let f(y) = marginal distribution for y
* If y is continuous 🡺 , if y is discrete = 
* Bayes’ rules related conditional density of f(y|x) to f(x|y) and f(y) (a joint related to another joint and a marginal)
* Special case = 2 sets A and B 🡺 
* Proof:
* Let X = indicator that event A has occurred. Let Y = indicator that event B has occurred
* Plug into discrete version of Bayes’ Rule
* Rules of probability Proof:
* 
* **Diagnostic tests** 🡪 Let + and – be events that the result of a diagnostic test is positive or negative
* **D** and **Dc** = events that the subject does and does not have a disease, respectively
* **Sensitivity** = probability the test is positive given the subject has the disease (TP) = **P(+|D)**
* **Specificity** = probability the test is negative given the subject does not have disease (TN) = **P(-|Dc)**
* **Positive predictive value** = probability the subject *has* the disease, given test is positive, **P(D|+)**
* **Negative predictive value** = probability subject does not have disease, given test is negative, **P(Dc|-)**
* **Prevalence** of disease =*marginal* probability of disease **P(D)**
* **Diagnostic likelihood ratio of a positive test**, **DLR+ = P(+|D) / P(+|Dc) = sensitivity / (1 – specificity)**
* **Diagnostic likelihood ratio of a negative test**, **DLR- = P(-|D) / P(-|Dc) = (1-sensitivity) / specificity**
* *Ex: Study comparing efficacy of HIV test reports on experiment that concluded that HIV antibody test have sensitivity = .997 and specificity = .985*
* *Suppose a subject from a population with prevalence of HIV = .1% receives a positive test.*
* *What is the probability the subject has HIV? 🡪 P(D|+)*
* Bayes’ Rule is convenient in these sorts of settings b/c, in principle, it’s easier to get sensitivity + specificity values by just taking blood samples from a set of people you *know* are HIV+ and seeing what proportion of them had the test comes up positive.
* Then, take a group of people you *know* to be HIV- and seeing the proportion that have a negative test result
* Could get these numbers or *estimates* of these numbers
* very simplistic treatment of how you‘d actually get a sensitivity + specificity
* there's lots of issues 🡺 how do you actually know if you're working in an area where the tests are difficult, how do you actually know if a person has the disease or not, or if you wait so long to where the disease is very clinically relevant, then are we evaluating the test in a stage of the disease where it's not interesting for when we’d be applying the disease?
* There's a lot of issues in development + evaluation of tests + constructing their validity
* Mathematically, we want P(D|+), *given* sensitivity **P(+|D) =** .997, specificity **P(-|Dc)** = .985, and prevalence P(D) = .001
* Use Bayes’ 🡪 
*  =  
* So, *in this specific population*, a positive test result suggests only a 6% probability the subject has the disease 🡺 **Positive predictive value = 6.2%**
* Seems awfully low, and this is due to the very low prevalence and somewhat modest specificity
* Suppose we knew the subject was an IV drug user + routinely had intercourse w/ an HIV-infected partner 🡺 higher **prior** that the subject has HIV, since prevalence among *that* population would be higher, so our **PPV** would be higher
* External info changed our original conclusion 🡪 actual test value does not chance, only the prevalence in the calculation changed
* We notice evidence implied by a positive test does *not* change due to prevalence of the disease in said subject’s population, only our *interpretation* of that evidence changes
* So what component of the calculation does NOT change, regardless of changes in prevalence? 🡺 **diagnostic likelihood ratios**
* Bayes’ Rule:  and  so  🡺 **DLR+**
* This = a ratio == **post-test odds of having disease = DLR+ \* pre-test odds of having disease**
* So we have some inkling of whether someone has the diseased based on prior knowledge before administrating the test, then we perform the test, which yields data, and DLR+ - the factor by which we multiply pre-test odds to get post-test odds
* So, from before whether a subject is a IV drug user w/ and HIV+ partner is irrelevant to the DLR+, which tells us odds of having disease are increased by x amount given a positive test, regardless of prevalence
* Whereas positive + negative predictive values (LHS of equation above) inherently factor in prevalence (P(D) and P(Dc)) 🡺 higher pre-test odds
* This is basically post-test odds = likelihood ratio (probability model and data combined) times the pretest odds
* This is a very appealing mirror to how we think the scientific process should work.
* i.e. Start out w/ an a-priori set of hypotheses (a hypothesis + its complement) + then collect data
* That **data informs your belief, +** now, we have a *post*-test odds of disease.
* If we were to run *another* test, the starting point for your new **prior** odds would be the post-test odds/**posterior** after this 1st test.
* It all works out just fine if you take 2 tests that’re both positive, so the diagnostic likelihood ratios just multiply.
* But, in terms of Bayesian thinking, this is the idea that we have some **prior**, we **update** it with **data**, we get a **posterior**, + now that posterior is the **new prior**.
* This also codifies a lot of scientific discussion 🡺 if a prior is *absolutely fixed* at a specific point, data is irrelevant, as nothing is going to move it.
* Suppose a subject has a positive test, and DLR+ = .997/(1-.985) = 66
* The result of having positive test == post-test of disease = 66X more than original pre-test odds, regardless of the prior and of the population
* i.e. "hypothesis of disease == 66X more likely (more supported by the data) than the hypothesis of no disease”
* Suppose a subject has a negative test, and DLR\_ = (1- .997/.985 = .003
* The result of having negative test == post-test odds of disease = .3X (.3% of) the original pre-test odds, regardless of the prior and of the population
* i.e. "hypothesis of disease == .003X that of the hypothesis of no disease, given a negative test”
* For a Bayesian, a probability is not objective, but instead is a **quantification of belief**
* Frequentist = person has disease or they don’t
* **NOTE** on the nature of actually collecting data to inform these calculations
* Usually very difficult to know, *conclusively*, whether or not someone has a disease when developing a test
* Very difficult to develop things like *actual, real* prevalence estimates that’re relevant to the person you’re talking about, w/ respect to the disease.
* Very difficult to have whatever samples you're using to develop sensitivity and specificity actually be indicative of the population of samples that the test will be applied to in actual clinical practice.
* So, even though these calculations are very simple + highlight Bayes' rule quite nicely, this is NOT all there is to the world of diagnostic testing and validation, which is a very, very deep subject that involves quite a bit more than Bayes' rule.

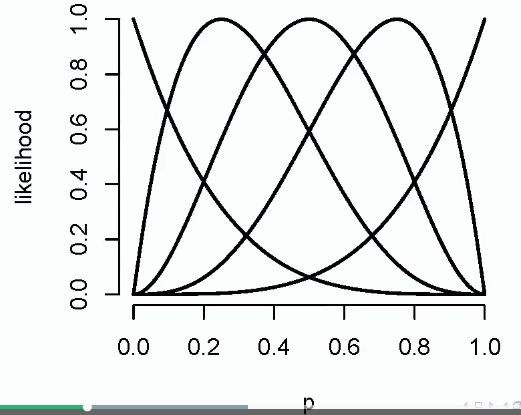
**Likelihood**

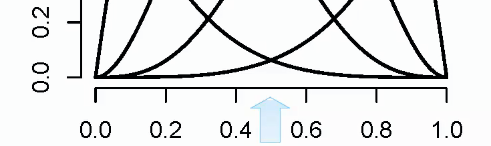
* **Likelihood =** mathematical construct used to relate data to a population
* **Maximum likelihood =** way of using likelihoods to create estimates
* Common + fruitful approach to stats = *assume that data arises from a family of distributions, indexed by a parameter that represents a useful summary of the distribution*
* So, **likelihoods** arise from a probability distribution (used to connect data to a population), and the likelihood of a collection = the **joint density** evaluated as a function of the parameters, w/ the data being fixed
* Likelihood analysis of data used likelihood to perform inference regarding the *unknown* parameter
* Ex: assume data come from Gaussian distribution 🡺 all we need = mean + variance (to get SD)
* Goal = use the data to infer the unknown parameters (mean + variance) w/ the idea that the mean + SD = population parameters (since Gaussian distribution = the model for the population) and the data = sample parameters we use to estimate unknown population parameters
* Given a statistical PMF or a density, we say *f(x,ϴ),* where x = a vector and ϴ = an unknown parameter, then we say the **likelihood** is *f* viewed as a function of ϴ for a fixed, observed value of x
* **Likelihood** has 3 important properties
* Ratios of likelihood values measure the **relative evidence** of one value of the unknown parameter relative to another
* Given a statistical/probability model + observed data, the **likelihood principle** states that all relevant info contained in the data regarding the unknown parameter is *contained in the likelihood*
* Likelihood principle has a mathematically-correct proof, but not all agree with it interpretation + applicability
* *Has very-far reaching consequences in stats 🡪 p-values, much of hypothesis testing, + other staples of stats become questionable if we take this point as true*
* Statistical model MUST be specified correctly, but we almost never always do, we just assume
* If {Xi} are independent random variable, then their likelihoods multiply (likelihood of all parameters given all Xi = the product of individual likelihoods)
* Ex: Flipping a coin w/ P(success) = ϴ
* Recall the **mass function** for 1 coin flip x 🡪 f(x,ϴ) =  where x is either Tails (0) or Heads (1)
* Suppose we get Heads, then the likelihood is 
* Therefore  == there’s twice as much *relative* evidence supporting the hypothesis that ϴ = ½ (coin is far) than the hypothesis that ϴ = .25 (biased coin towards Tails)
* Extended Ex: Flipping a coin 4 times w/ P(success) = ϴ and w/ results = {1,0,1,1}
* ***L(ϴ,1,0,1,1}*** *🡺 L is a function of theta depending on the data we actually observed*
*   (1st flip\*2nd flip x 3rd flip\*4th flip) 
* This likelihood *only* depends on the *total* number of H and the *total* number of T, not on order such as written in shorthand as 
* *Order doesn’t matter* (a property of likelihoods) 🡺 all relevant info about parameter ϴ is contained just in the fact that we got a specific # of H and specific # of tails
* This shorthand lets us know we only need the totals for H and total flips (from which we can get totals for T), not when they occurred (i.e. only knowing we have 1 T, 3 H = a **sufficient statistic**)
* Consider  == there’s more than five times as much *relative* evidence supporting the hypothesis that ϴ = ½ (coin is far) than the hypothesis that ϴ = .25 (biased coin towards Tails)
* What we’d like is to consider likelihood ratios of all values of the parameter ϴ between 0-1
* Constants that don’t depend on ϴ don’t matter in likelihood b/c when we take the ratio and this constant is in the numerator + denominator, it just cancels out
* *So, the likelihood’s interpretation should be* ***invariant*** *to constants that are not a function of the parameter*
* i.e. *raw absolute value of the likelihood isn’t all that informative*
* A likelihood plot displays ϴ by *L(*ϴ,x), and it’s usually divided by its max value such that its height = 1
* B/c likelihood measures *relative* evidence, dividing the curve by its max value (or any value in that matter) does NOT chance its interpretation
* Likelihood plot for the 4 coin flips above:  🡪 height = 1 as a max/peak, values as a function of ϴ
* As we leave peak height = 1, corresponding values of ϴ have less and less supporting evidence
* The peak likelihood value @ which we divided all likelihood values by = *the best-supported point in the data ==* **maximum likelihood estimate point**
* No matter what we divide it by, we get a likelihood ratio > 1
* We’d interpret this plot in such a manner:
* Take points of ϴ = .4 and ϴ = .6, get their heights, and the ratio of these heights = relative evidence for hypothesis supporting ϴ = .4 or ϴ = .6
* B/c we divided by the maximum likelihood, every value we look at is the relative evidence of that specific ϴ value when compared to the maximum likelihood (point that is best supported by the data)
* ϴ = .5 has a likelihood value of ~ .593
* The horizontal line at likelihood = 1/8 🡺 every point that falls between the endpoint of this line is such that there’s no other point that more than 8x more supported
* 
* The points where the curve meets the line is exactly == 1/8, which means that point is exactly 8x worse supported, given the data, than .75 (3 H out of 4 trials), the maximum likelihood value
* Taking any point w/in the endpoints of this line means we can’t find another points that’s more than 8x more supported
* Ex: .4 has likelihood of .364, but its ratio relative to the maximum is < 1/8, so its ratio w/ everything else within the interval is less than 8x worse
* This means we’re not going to be able to find, for ϴ = .4, another point anywhere on the curve that’s more than 8x better supported than its likelihood
* The collection of data values that lie *on* the horizontal in between the points where it intersects the likelihood curve are well-supported values.
* As we move the line upwards, fewer points stay in the interval
* **Maximum Likelihood Estimate (MLE) =**  **= argument maximum over** ϴ of the likelihood having plugged in the data, x
* Another interpretation = value of ϴ that makes the data we observe most probable
* Ex: if we indeed got 3 H in 4 flips, the MLE = success probability of the coin that would make this observed data most probable
* MLE for ϴ is *always* the proportion of H in coin flips (for IID coin flips)
* Proof:
* Let x = # of heads, n = # of Bernoulli trials
* Recall:  = ϴ to # of heads, 1-ϴ to # of tails
* It’s easier to *maximize* the **log-likelihood** = 
* Almost a general principle in stats 🡪 when we have a bunch of independent things + when we want to maximize the likelihood, we’re better off maximizing the log-likelihood
* Maximizing log = maximizing the functions b/c log = increasing monotonic function
* Also, multiplying a bunch of independent things means we’ve gotten the joint density or mass function
* multiplying a bunch of independent things raises them to some power, which are complicated to work with, and addition is much easier to work w/
* log convert products into sums, which is useful (x turns into a coefficient to log(ϴ)
* Take derivative 🡺 
* Maximize by setting = 0 🡪 implies , which is solved at ϴ = x/n
* NOTE: Second derivative , provided x <> 0 or n (all failures or all successes)
* **Treat likelihood as the arbiter of evidence and likelihood ratios as measurements of evidence**
* What constitutes “strong” evidence?
* Ex: repeatedly flip a coin + entertain 3 hypothesis, H1: ϴ = 0, H2: ϴ = .5, H3: ϴ = 1
* 
* P(X | H1) = P(heads | coin is tails on both sides)
* P(X | H2) = P(heads | coin is fair)
* P(X | H3) = P(heads | coin is heads on both sides)
* Getting H on 1st flip = likelihood ratio of 0 for supporting the 2-tail hypothesis H1 vs. the coin is fair w/ 2x as much relative evidence supporting the coin is 2-headed vs. the coin is fair
* 2 is not very strong evidence, considering we have only 1 flip to go off of
* 2 H in a row increases evidence for H3, even more-so for 3 H
* Using this example as a guide, researchers tend to think of likelihood ratios of:
* 8 = moderate evidence 16 = moderately strong evidence 32 = strong evidence

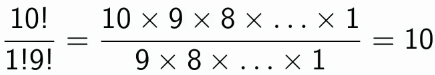
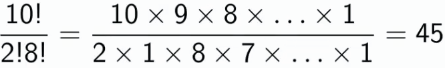
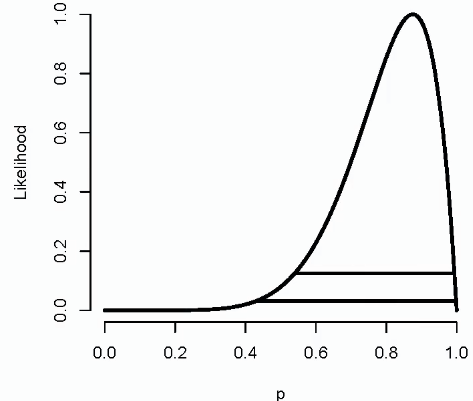
*of one ratio over another*

* B/c of this, it’s common to draw reference lines at these values on likelihood plots
* Parameter values > the 1/8 reference line are such that no other point is more than 8x more supported, given the data
* Consequences Of Kind Of Adopting This Style Of Analysis.
* Pretty much every major paradigm in statistics (Bayesian, frequentist, this likelihood paradigm) agrees that if you assume some probability model + act as if it's true, then the **likelihood ratio** is a central component to the theory.
* If you take enough mathematical statistics, you'll see this.
* The particular paradigm above then goes *beyond* this relatively benign use of likelihood ratios that occur in other areas.
* The above is not just saying that the likelihood ratio is useful but that **likelihood ratios measure relative evidence, + that given a statistical model + observed data, all the relevant info is contained in the likelihood**
* This has far reaching consequences to the field of stats.
* If we go beyond just saying “likelihoods are useful” to saying “not only are the useful but they have these properties”, it changes quite a bit of statistics.
* Much of stats is devoted to things like hypothesis testing + p-values + other variants of statistics where the interpretation of the statistics involves potentially fictitious repetitions of an experiment
* Ex: interpretation of a CI is quite confusing, but it's something along the lines of “if you were to use this technique over and over again, you’d obtain these intervals that contain the things they were trying to estimate 95% of the time”.
* If you adopt this strong variant of interpreting likelihoods, it that suggests that such an interpretation *can't be valid b/c it involves potentially fictitious repetitions of the experiment which do not depend on the likelihood for the data at hand*, so it cannot possibly be useful or cannot have any additional evidence.
* So, some things get disputed if you adopt this paradigm (p-values, hypothesis testing, multiple corrections)
* This is very disputed b/c, in many ways, these techniques seem very central to the idea of statistics.
* Regardless of what kind of paradigm statistics you're in, **higher likelihoods** **generally** **refer to better supportive values of the parameter**

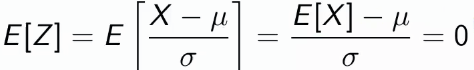
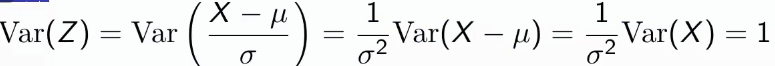
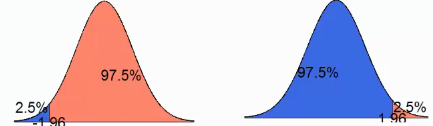
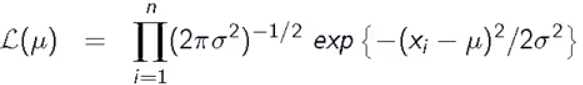
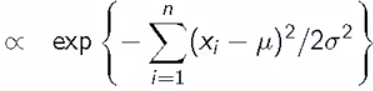
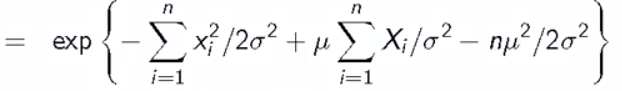
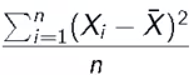
**Bernoulli Distribution and Binomial Trials**

* **Bernoulli distribution** arises as a result of a binary outcome, and **Bernoulli random variables** take *only* the values 1 and 0 w/ probabilities p and (1-p), respectively
* **PMF for Bernoulli random variable X = P(X=x) =** 
* **Mean** of Bernoulli random variable = p, **Variance** of Bernoulli random variable = p(1-p)
* If X = Bernoulli random variable, we typically say X=1 is “success”, X=0 is “failure”
* **Bernoulli Likelihood function =** If we have several IID Bernoulli observations, say x1,….,xn, then the likelihood =  🡺 *only depend on the sum of the xi*
* B/c n is fixed + assumed to be known, this implies that **sample proportion of successes**  contains all the relevant info about p (insofar as the likelihood codifies it)
* Can maximize the Bernoulli Likelihood *over p* to obtain  == the **MLE estimator** for p
* *This all depends that we’ve correctly modeled the data as IID Bernoulli*
* Ex: 4 coin flips 🡪 4 possible outcomes = 1H, 2H, 3H, 4H
* 4 possible likelihoods from this experiment🡺  (normalized to height = 1)
* MLE for 0H = 1 or 0, MLE for 1H = .25, 2H = .5, 3H = .75, 4H = 1 or 0
* For 4H or 0H, the likelihood is not *entirely* 1 or 0, there’s substantial uncertainty
* The likelihood is correctly codifying the info that it is possible, even if the coin is fair, to get 4 consecutive T

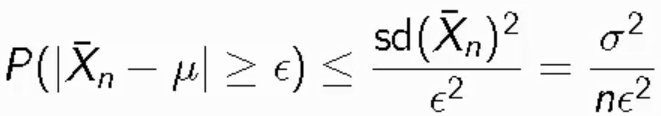


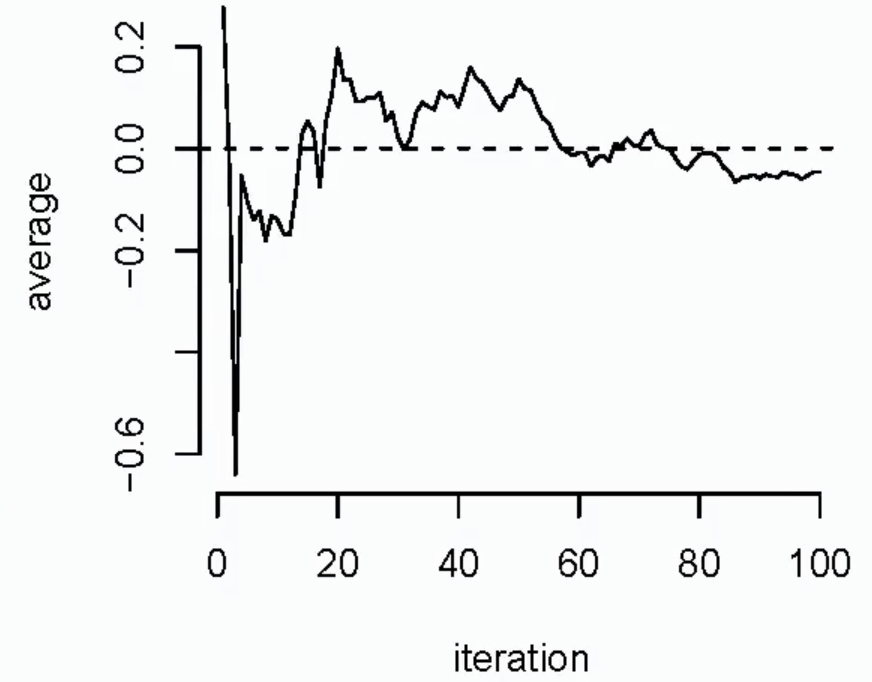
* It's just much less likely than if the coin was unfair towards some value closer to 0.
* So, likelihood is not entirely shoved up against the vertical line at zero, it just gets closer + closer to that vertical line as you continue to flip and get tail after tail after tail.
* **Binomial random variables** are based on Bernoulli, + are obtained as the **sum of IID Bernoulli trials**
* Key variable from a Bernoulli experiment = # of H, so why not create a random variable that is the # of H 🡺 **Let X1, ..., Xn be IID Bernoulli(p), then X =  = a binomial random variable**
* **Binomial Mass Function (PMF) =**  for x = 0 (every flip is T), …, n (every flip is H)
* “probability X takes any specific value x is “n choose x” times …..
* Remember  🡺 counts # of ways to select x items out of n total w/out replacement, ignoring order
* 
* Choosing 1 tie out of 10 ties = 
* Choosing 2 tie out of 10 ties = 
* In general: 
* Justification of the binomial likelihood
* P(6H in 10 flips w/ P(s) = p) 🡪 P(6H, 4T) in *any* order =  and there are  possible pairings of 6H, 4T
* Binomial Mass Function Ex:
* Friend has 8 children, 7 of which are girls, w/ no twins
* If each gender truly has 50% probability in each birth, what is P(7+ girls in 8 births)?
*  🡺 *add probability of both 7 and 8 girls*
* This calculation = example of a **p-value** = probability, under a null hypothesis, H0, of getting a results *as extreme* or *more extreme* than the one observed/obtained
* We have evidence here, and we think having 7/8 girls is odd, so maybe 50% chance of boy or girl is off for this particular family
* This p-value is calculating the probability of getting/seeing this results, if our null that we have 50% chance of boy or girl is true
* If low, our null is probably *not* right
* Likelihood for this experiment:  🡺 see p=.5 is in the 1/16 likelihood, but not the 1/8 likelihood
* The likelihood is peaking around p=7/8, + the curvature sort of gives a sense of the relative evidence for the collection of possible values of p.

**The Normal Distribution**

* A random variable follows a **normal/Gaussian distribution** w/ 2 parameters (shift + scale) of mean μ and variance δ2 if the associated density is =  🡪 centered at μ + δ2 says how flat or peaked it is
* If random variable X has this density, then **E[X] = μ** and **Var(X) = δ2**
* To say random variable X is modeled by/follows the Gaussian, write **X ~ N(μ,δ2)**
* When μ = 0 and δ2 = 1, we have **the Standard Normal Distribution,** whose density function = 
* Standard Normal random variables often labeled as ***Z***
* ~68% of the normal distribution lies w/in 1 SD of the mean, ~95% w/in 2 SD, ~99% w/in 3 SDs
* Can get from a nonstandard normal to a standard normal very easily
* **Standardizing** = If **X ~ N(μ,δ2)**, then  == the standard normal
* works for ANY distribution 🡪 take any random variable X, subtract the population mean, and divide by its SD = results in a random variable w/ mean = 0 and variance = 1
* and if X was normal, then resulting Z is normal
* **WE can then check that Z has the correct mean and variance**
* 
* Factor out 1/δ, and E[X] for normal = μ, so μ – μ = 0
* 
* Factor out 1/δ and square
* ***Remember this rule:*** if we **shift a variable** by a constant (i.e. subtracting μ), **variance doesn’t change**
* Var(X) = δ2, so we get 1
* Reverse = If Z is the standard normal**,** we can get a non-standard normal X via
* scaling Z by δ2 , shifting by μ
* The **non-standard normal density** = 
* **This** is obtained as a “plugging into” the standard normal density
* Take  and instead of plugging in Z, plug in , then divide all by δ, we get the non-standard normal density
* μ = shift parameter, shifting the distribution right or left
* δ2 = scale factor
* whenever we take some kernel density and create a new family by plugging in  + dividing all by δ, we get a new family of densities w/ mean = μ and variance = δ2
* So this is an interesting way of taking a root density with mean = 0, variance = 1, and creating a whole *family* of densitiesw/ mean = μ, variance = δ2 **(location scale families)**
* More facts about normal density
* 68-95-99% of the normal density lies between 1, 2, 3 SD’s of the mean, respectively, symmetric about μ  (34% above and below mean for 1 SD)
* **-1.28, -1.645, .196, -2.33** are the **10th, 5th, 2.5th, 1st percentiles** of the standard normal distribution, respectively
* By symmetry, **1.28, 1.645, 1.196, 2.33** are the **90th, 95th, 97.5th, 99th percentiles** of the standard normal distribution, respectively
* 
* Between +/- 1.96, we get 95% of the data, so we only have a 5% chance (2.5% of each tail) of lying outside of that range of SD’s from the mean (used for **confidence intervals** and **tests**)
* Ex: what is the 95th percentile of a **N(μ,δ2)** distribution?
* We want to find the point x0 such that P(X < x0) = .95 (*at what value of the distribution are 95% of the data points below it?*)
*  **🡺 normalize 🡺**  **🡺**  **= .95**
* Therefore, , or, solving 🡺 **x0 = μ + 1.645\*δ**
* In general, **x0 = μ + z0\*δ** where z0 = appropriate standard normal quantile
* In R, just do **qnorm(.5,mean,variance)**
* Ex: What is the probability a random variable from a **N(μ,δ2)** distribution is 2SD above the mean?
* We want to find the point x0 such that P(X < x0) = .95 (*at what value of the distribution are 95% of the data points below it?*)
*  **🡺 normalize 🡺**  == probability a standard normal random variable is > 2 🡺` 
* To know the probability that a random variable is bigger or smaller than any specific # or between any two #’s, just take those #’s + convert them into standard deviations from the mean (could be fractional, or floating like 1.12 SD’s from the mean, etc.)
* This is done by subtracting off μ and dividing by **δ** + then reverting that calculation to a standard normal calculation
* To know the probability a random variable is bigger than 3.1 (probability of a child being taller than 3.1 feet), we’d need is the population mean μ and the standard deviation **δ**, then take 3.1, subtract μ, divide by **δ**.
* This converted quantity = 3.1, in feet, to *SD units*.
* Then, we just do the remainder of the calculation using the standard normal.
* R code can do these calculations very quickly but it's worth doing by hand to get used to working with densities + to get used to what these calculations refer to
* More properties of the normal distribution:
* Normal distribution is symmetric + is peaked about its mean μ, and therefore μ = median = mode
* **\*\*\*\*Linear Transformations of Normal Random Variables\*\*\*\* =** <https://onlinecourses.science.psu.edu/stat461/node/74>
* **A constant times a normally-distributed random variable is *also* normally distributed**
* 
* 
* **Sums of (*jointly*) normally-distributed random variables are *also* normally distributed, even if variables are dependent**
*  `
*  🡺 
* *if we add 2 normally-distributed random variables together, they‘ll end up being distributed as a normally distributed random variable w/ a mean = sum of the individual means, + a variance = sum of the individual variances*
* **Sample means of normally-distributed random variables are *also* normally distributed**
* *They’re estimating the normal population mean and variance*
* The square of a standard normal random variable follows the **chi-square distribution**
* The exponent of a normally-distributed random variable (eX) follows the **log-normal distribution**
* This means “take the log of something + *then* it becomes normal”, NOT the log of a normal random variable
* Can’t do this b/c there’s a non-zero probability the random variable is negative and we can’t take the log of a negative
* Many random variables, which are *properly normalized*, will *limit the normal distribution*
* It also turns out, if you have IID observations + properly normalized sample means, their distribution will look like a Gaussian distribution (not entirely but close), regardless of the underlying distribution the data comes from.
* Ex: Roll a die + look @ the distribution of a die roll, which doesn't look very Gaussian but looks more like a uniform distribution for values 1-6.
* Now, take a die, roll it 10 times, take the *average*, + then repeat *that* process over + over again + think about the distribution of these averages of 10 die rolls.
* It turns out it'll look quite Gaussian
* At any rate, that's the rule 🡺 That random variables, *properly normalized*, (w/ some extra conditions) will limit to a normal distribution.
* That's how the Gaussian distribution became a “Swiss army knife” of distributions 🡺 pretty much anything you can relate back to a *mean of independent things* tends to look Normal-ish in a distribution.
* Mathematically/formally, if they're IID + you normalize the mean in the correct way, then, you get *exactly* the standard normal distribution.
* That is an incredibly useful + a very historically important result = **the central limit theorem**
* Ex: If Xi are IID and ~ N(**μ,δ2**) w/ a known variance, then what is the likelihood for/associated w/ **μ?**
*  🡺 take product of likelihoods for each individual Xi (observations)
* Move the above product into the exponent 🡺
* assuming **δ2** is known, so  divvided by n-2 can be thrown out b/c likelihood doesn’t care about factors/proportionality that *don’t depend on* **μ,** since it’s the parameter of interest
*  == “proportioned to”, since we dropped out things NOT relating to **μ**
* Factor out the square 🡺
* Again, 1st summation term *doesn’t depend on* **μ** 🡪 remove = 
* Replace summation to n with x\_bar, since x\_bar\*n = the summation anyway
* So, what is the MLE for μ when δ2 is known?
* WE calculated likelihood for μ =  but exponents are annoying, so take the log 🡺 , then derive w/ respect to μ =  and solve for when this = 0, *which is clearly when x\_ = μ*
* This yields the fact that **x\_ the is the MLE of μ**
* **So, if data is normally distributed, our estimate of a population mean μ is the sample mean x\_**
* Also, since the calculation *doesn’t* depend on **δ2**, this is *also* the MLE of μ when ISN’T **δ2** known
* The MLE for **δ2** (rather than μ) 🡺 , which == the **biased version (divided by n, not n-1) of sample variance**
* The MLE for δ is just the square root of the above
* To do FULL likelihood inference, the bivariate likelihood of (**μ,δ2**) is difficult to visualize
* It’s a surface w/ **μ** and **δ2**on different axes and the likelihood on a vertical axis (**likelihood surface** rather than likelihood function)
* There *are* methods for getting rid of μ and looking at δ2 and for getting rid of δ2 and looking at μ

**Asymptotics (Limits and Law of Large Numbers)**

* Reminder of **limit** = for any fixed distance, we can find a point in the sequence such that the sequence if closer to the limit than that distance from that point on
* Ex: a1 = .9, a2 = .99, a3 = .99 🡺 clearly this sequence **converges** to 1
* |an – 1| = 10-n 🡺 where n = the point in the sequence
* If we take any distance, ε, we can find an n such that is 10-n smaller than ε, and b/c 10-n keeps getting smaller as n grows, it’s 10-n is smaller than ε form that point onward
* This problem is harder for random variables
* Consider  == sample average of 1st n of a collection of IID observations
* Ex:  could be the average of n coin flips (sample proportion of H)
* We say  **converges in probability** to some limit if, for any fixed distance ε, the *probability* of being closer (or farther away) than said distance to the limit converges to 1 (or zero)
* P(| - limit| < ε) 🡪 1 for any fixed distance ε
* Establishing that a random sequence of variables actually converges = difficult
* To make it easier, there is the **Law of Large Numbers (LLN)** theorem= if Xi, …, Xn are IID from a population with mean μ and variance at δ2, thenconverges *in probability* to μ
* **Averages of IID observations/random variables X converge to μ of the populate from which the observations were drawn**
* This basically says if we collect an infinite amount of data, we get the # we want to estimate = μ
* There are many variances of LLN, since there are many distributions where δ2 is unknown, this is the “lazy” one
* If willing to make assumptions that X’s all have a finite variance, we can prove LLN using Chebyshev’s inequality == probability a random variable X > k SD’s from the mean is < 1/k2
* Therefore, for the sample mean, this is 
* Pick any distance ε and set  to get 
* b/c to establish convergence of a limit of numbers we need an ε
* b/c to establish convergence in probability, we *need to show that the probability that  being > ε converges to 0 OR being < ε converges to 1* (equivalent statements)
* k is NOT a random variable, since ε is a value we picked and while  *is* a random variable, its SD is δ2/sqrt(n) and is just a number
* So, as n goes to Inf, neither δ2 nor ε 2 are changing, so the whole RHS of the above inequality heads to 0
* So, the probability the average of the random variable is more than ε away from the mean μ () goes to 0 as n goes to Infinity (or goes to 1 if looking for probability of being *less* than ε away from the mean μ
* Either statement is saying that  converges in probability to μ as n grows
* Ex: Averages of cumulative sample means of simulated random normals w/ μ = 0



**CODE**

**num\_sim <- 10000**

**# generate list of 10k averages of increasing large samples from normal distribution**

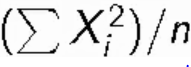
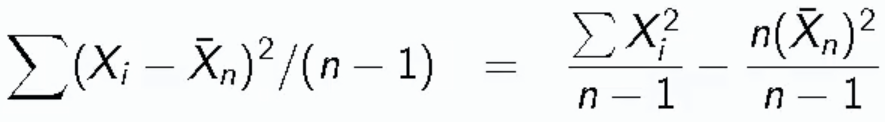
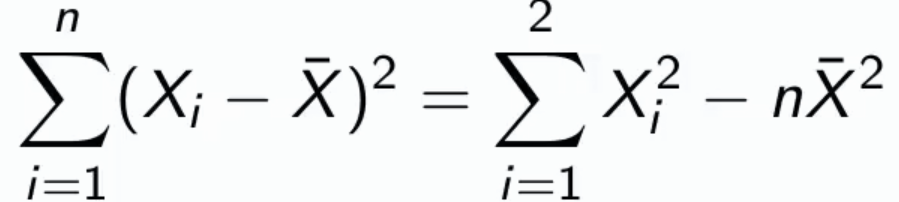
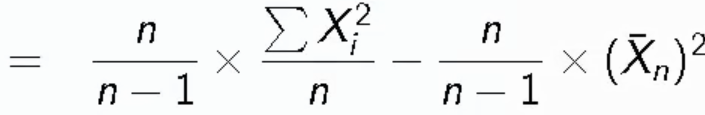
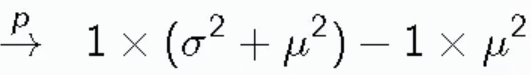
**x <- cumsum(rnorm(num\_sim)) / 1:num\_sim**

**plot(x = 1:num\_sim, y = x, type = "l", xlab = "Iteration",**

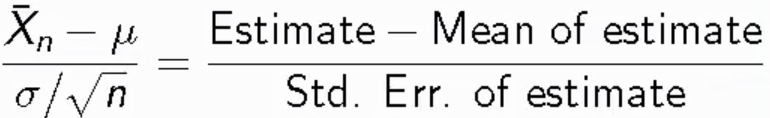
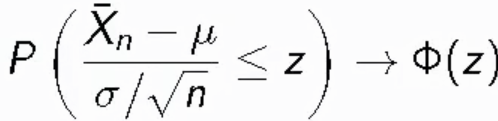
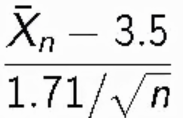
**ylab = "Average", frame = F)**

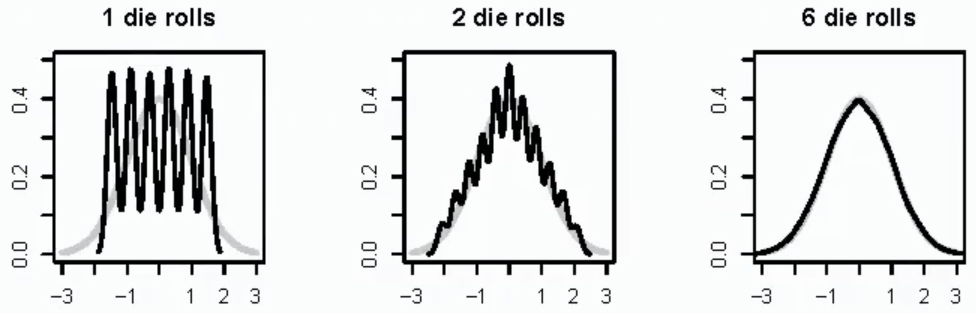
**# add mu = 0 as the limit reference line**

**abline(h = 0, lty = 2)**

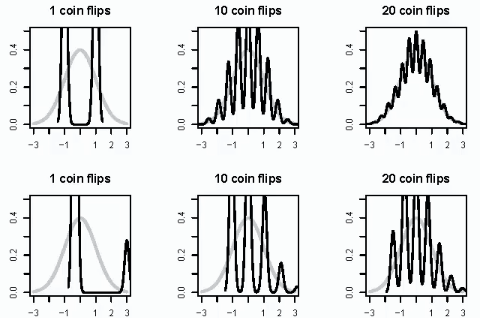
* Dashed line = asymptote
* See as we add more random variables (iterations), the average of more and more random variables converges to μ = 0
* Fact about LLN
* Functions of convergent random sequences converge to the same function evaluated at the limit
* Includes sums, differences, products, etc.
* Ex: ()2 converges to μ2
* This is different than  (average of squared observations), which converges to 
* If we sum up the squares of *individual* observations/random variables Xi, it no longer converges to μ2
* *Square of the average vs. average of the squares*
* We can use this to **prove the sample variance δ2**
* 
* LHS = unbiased sample average
* RHS 🡪 remember shortcut for the numerator == 
* Rearrange terms + multiply and divide by some n’s 🡺 
* Note: if multiply convergent sequences, they converge to the product of the limits (
* same for adding and subtracting, etc.
* n/(n-1) clearly converges to 1, we know converges to  and that  converges to μ2 🡺  = δ2
* Hence, we also know that the sample SD converges to δ
* So we have out LLN and w/ a couple rules, we know:
* An IID **estimator** is **consistent** if it converges to what we want it to estimate (of the population)
* LLN states that the sample mean is consistent
* It was shown that sample variance and sample SD are consistent as well
* Also recall the sample mean + sample variance are *unbiased* (sample SD IS biased)
* *NOTE: Consistency is a very weak property* for an estimator to be useful
* “Estimator is consistent” is not even a necessary or sufficient property for an estimator to be useful
* If something converges to μ + ε where ε = some miniscule # that is of no importance, then that estimator is “not consistent”
* Also, being unbiased is neither necessary nor sufficient for an estimator to be useful
* For example, Bias-Variance trade off 🡺 Estimators *can* be slightly biased, which can be desired b/c we improve on the variance.
* These are really a collection of properties that *describe* estimators, + we need to think about the collection of properties as a whole in order to evaluate an estimator.
* These various mathematical concepts are *useful* but never, in isolation, tell the full story on the utility of an estimator.
* Need to take properties as a collection to try + decide which estimators are the right ones to use
* Might be useful for eliminating dumb things (if definitely not consistent in a way that it doesn't converge anywhere near the estimate == probably not something to use)
* Ex: Estimator that's consistent but not very good
* Take only 1st half of collected observations == /2 rather than 
* That estimator is, of course, consistent as n goes to infinity 🡪 has fewer observations than if you took all of them, but still n is going to infinity in this case, it’s just n/2
* But, it's an obvious there’s better estimator right in front of you 🡺 the estimator using ALL data
* Here you have to account for the fact that the estimate w/ ALL data has a *lower variance* than the estimate w/ ½ of the data.

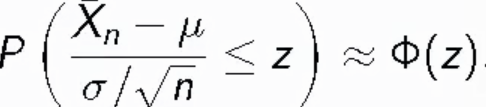
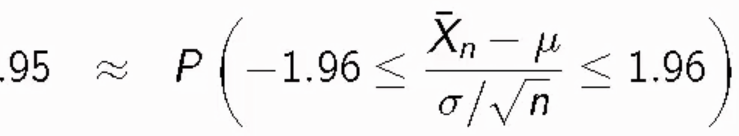
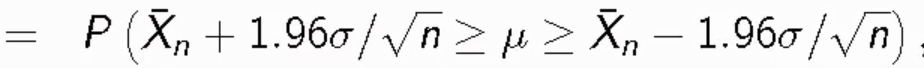
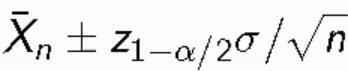
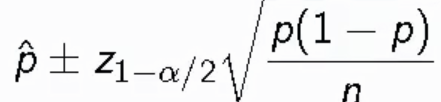
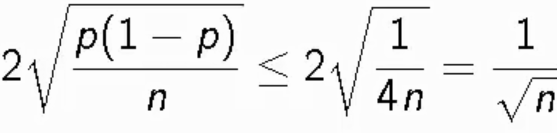
**CLT and CI’s**

* CLT = gives us a way to perform inference w/ random variables when we don’t know which distribution they come from + it applies in an endless variety of settings
* For this class’ purpose, the **CLT** = **the distribution of averages of IID random variables, *properly normalized*, becomes that of a standard normal/Gaussian distribution as sample size increases**
* In, the majority of inference done via statistical software (calculate a p-value or CI) probably the underlying motivation of what is done is asymptotics **(a method of describing limiting behavior**) + some use of the CLT especially if you have larger sample sizes.
* **CLT =** Let X1, …, Xn be a collection of IID random variables w/ mean = μ, variance = δ2
* Let  = sample average
* Notice the form of the **distribution function for the normalized quantity (i.e. for the normalized mean) 🡺** 
* It’s practically guaranteed that for any statistical estimate based on IID data, if we subtract off the population mean and divide by its SE, the result is most likely going to wind up limiting to having a standard normal distribution
* *This is our Z random variable μ = 0, variance = δ2 = 1*
* Then  🡪 Probability of the above Z random variable being <= a specific point z = **the standard limits to the standard normal distribution evaluated at that point z 🡺** 
* This is saying that probabilities associated w/ sample means look like probabilities associated w/ normals
* If we **standardize** the sample mean (i.e. make μ = 0, variance = δ2 = 1), the probabilities look like standard normal probabilities
* Ex: Simulate a standard normal random variable by rolling n, 6-sided die
* Let Xi = outcome for die i, and *remember* that μ = E[Xi] = 3.5 (*prior lectures*) and Var(Xi) = 2.92, so SE =  = 
* So **standardized mean** = 
* Histograms of Simulations 🡺 rolled the die, standardized the average w/ the formula above to center distribution @ 0 w/ variance = 1, w/ standard normal density in back 🡺



* Ex2: Let Xi = outcome (0/1) for the ith flip of a possibly unfair coin where p = true success probability
* Sample proportion = p^ = the average of coin flips (% of 1’s)
* *Remember* 🡺 **E[Xi] = p** (true success probability) and **Var(Xi) = p(1-p)**
* SEmean = , then  will be approximately normally distributed



* These normalized distribution of sample averages is more discrete here , compared to the normal and to the die example
* These distributions would converge to normality more quickly if the coin was fair (top row) than if unfair (bottom row)
* This is a problem w/ the CLT 🡪 doesn’t tell *how fast* distributions of normalized averages of IID random variables converges to the Gaussian/to normality, just that they does eventually (could be 10 samples, or 10000)
* CLT is used in practice b/c it’s very useful as an **approximation:**  **== normalized means have distributions the are approximately normal/Gaussian**
* Ex: Recall 1.96 = good approximation to the .975th quantile and -1.96 = good approximation to the .025th quantile of the standard normal
* CLT says ~.95 (95%) probability a standardized mean lies w/in +/- 2 SD, or just +/-2 for the normal
* Consider:  🡺 rearrange: = .95
* Therefore, according to the CLT, we can say that the random interval ** +/- 2 standard errors** contains μ (the non-random entity) with ~95% probability == **the 95% CI for μ**
*  contains **μ** with ~95% probability, where  == the  quantile of the standard normal distribution
* This CI is a mainstay of frequentstet statistics, and it’s basically an estimate for **μ** that accounts for uncertainty since we’re dealing with data that we’re treating as random
* Just giving the sample mean  has no acknowledgment of the random variation that we’re not accounting for
* If we take  + assume data are IID w/ a finite variance, we can apply CLT, which says if the distribution is cooperating + *n* is large enough, CI calculated contains **μ** w/ ~95% probability
* ***CI’s ARE VERY HARD TO INTERPRET*** (as a result of frequentist inference)
* The 2 numbers that make up the CI either contain **μ or not**
* Standard Frequentist logic = probability the CI contains μ or not is 1 or 0, no in between
* **So the “real” interpretation of the CI is that the procedure above repeated again and again would produce CI’s that contain μ 95% of the time**
* I.e. **CI procedure, given the CLT + other assumptions apply, creates intervals that, if we were to repeat the CI procedure over and over and over, ~95% of the time the CI’s will contain the value we’re trying to estimate**
* Shortcut 🡺 “**We are 95% confident that the interval contains μ” ==** enough hedging to count as a legitimate instance of the strict CI definition above
* ***Do NOT say “there’s a 95% chance”***
* **Slutsky’s theorem** allows us to replace the unknown δ with *s* (sample SD)
* Simple instance of the CLT: Sample proportions
* In the even that each Xi = 0 or 1 w/ common success probability p, then δ2 = p(1-p)
* The interval takes the form 
* Replacing p w/ p^ in our standard error results in a **Wad CI for p**
* *Since we’re esitmating p and therefore can’t use it in our formula*
* Note that the worst/biggest p(1-p) can be is if p = ½, so *p(1-p) <= ¼ so long as 0 <= p <= 1*
* Let **α** = .05 such that  = 1.96 ~= 2, then 
* Therefore,  is a quick CI estimate for p