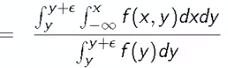
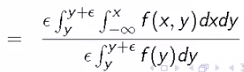
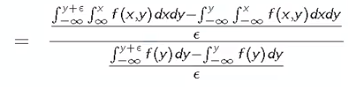
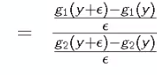
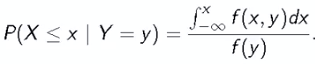
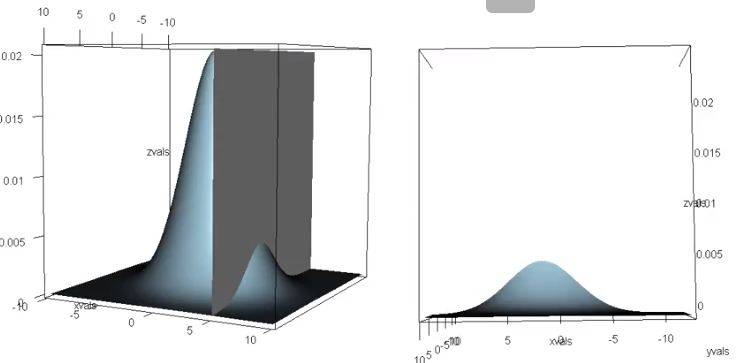
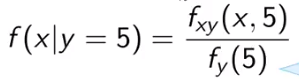
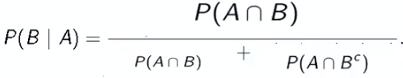
**Conditional Probabilities and Densities**

* 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

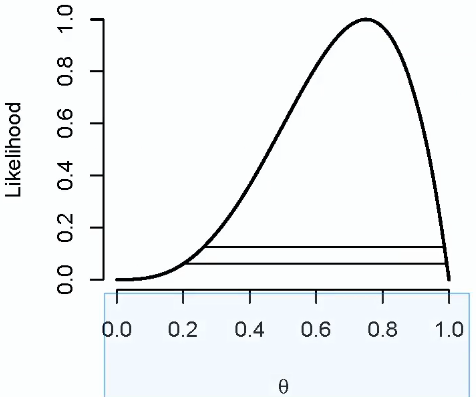
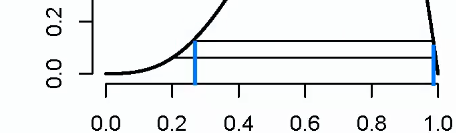
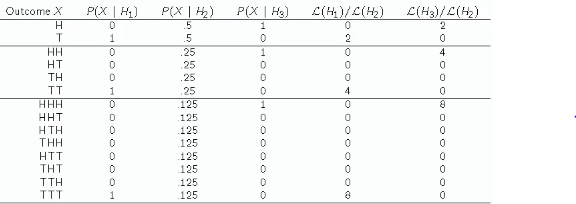
 🡺 

* 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 =** integratal 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 chance, 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**