***COURSERA: STATS W/ R SPECIALIZATION***

***COURSE 4 – Bayesian Stats***

**WEEK 1 – The Basics of Bayesian Statistics**

***1.1 Bayes' Rule***

**Conditional Probabilities and Bayes' Rule**

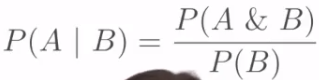
* 2015 Gallup Poll asked adult Americans whether they have personally ever used an online dating site, such as Match.com, eHarmony, or OKCupid.
* See distribution of responses by age group in a contingency table.



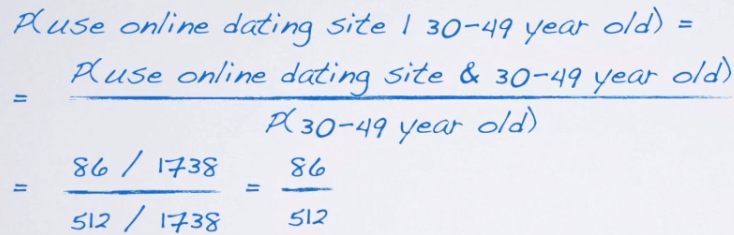
* Using these data, calculate what % of 30-49-year-olds use online dating sites  86/512 = ~17%.
* Formally, this = probability of using online dating sites **given 30-49 years old**.

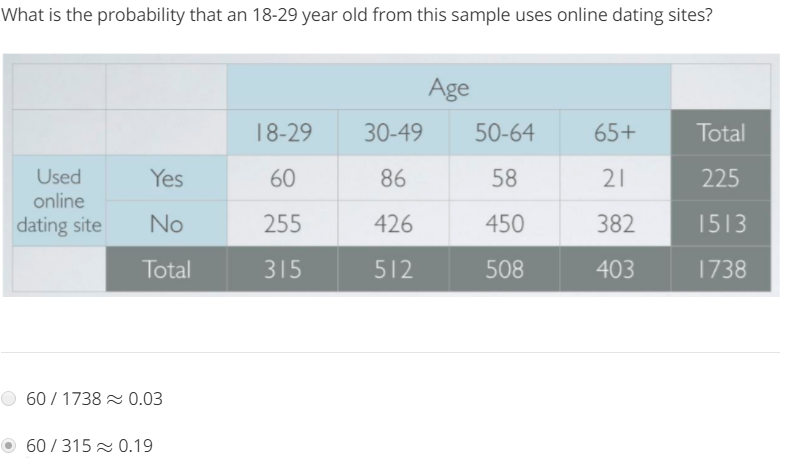


* Calculated this probability simply as a ratio of 2 frequencies, but we can formalize things a bit more
* Let event A = using an online dating site, event B = being 30-49 years old.



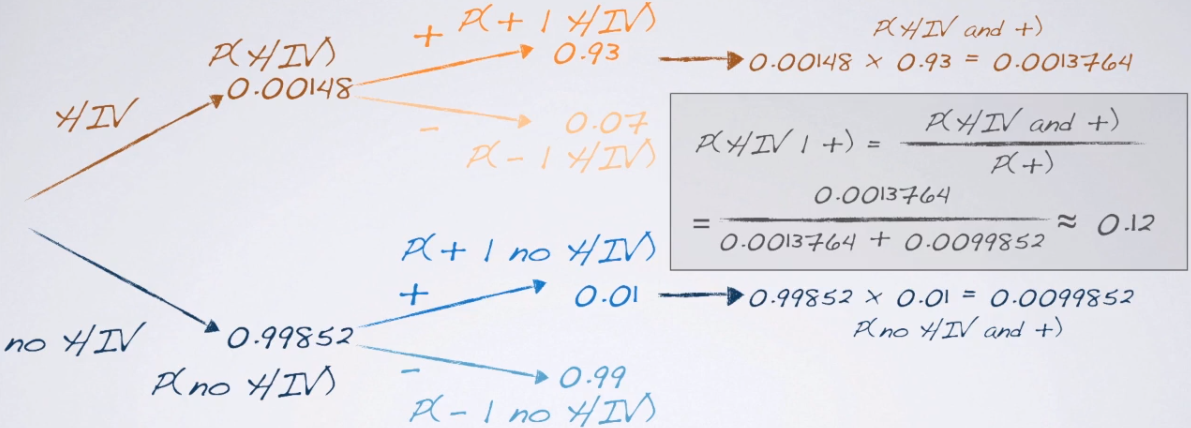
* Numerator frequency = # of times events A + B happened at the same time
* Denominator = # of times event B happened.
* Probability of A given B = probability of A + B divided by probability of B = **Bayes' rule**
* Thomas Bayes (1702-1761) = mathematician who established a mathematical basis for **probability inference** = means of calculating probability an event will occur in future trials from the # of times an event has *not* occurred
* Wrote his findings on probability in An Essay Towards Solving a Problem in the Doctrine of Chances published in 1763 after his death.
* Bayes' contributions = immortalized via fundamental proposition in probability = **Bayes' rule**
* Probability of using an online dating site, given being 30-49-year-old = **joint probability** of these 2 events divided by the probability of the event we're **conditioning** on.
* **Joint probability** = 86 / 1738 + **marginal probability** of being 30-49-year-old = 512 / 1738.
* A little bit of simplification, + we're again left w/ 86 / 512 = ~17%.

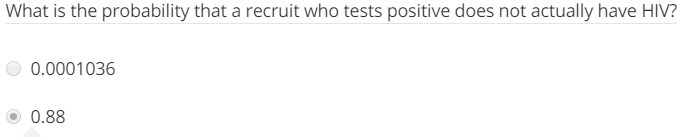


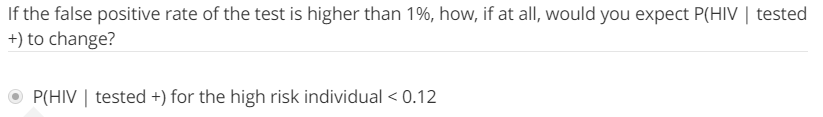


**Bayes' Rule and Diagnostic Testing**

* In early 80's, human immunodeficiency virus (HIV) had just been discovered + was recognized as a rapidly expending health epidemic.
* False negative (FN) result for communicable disease = very important personal + public health concern
* Specifically for HIV in the U.S. at the time, the safety of blood supply was a major issue.
* However, false positives (FP) also carried a lot of weight at the time due to stigma associated w/ testing positive for HIV, as well as complete lack of treatment options.
* In mid-80's, an HIV diagnosis = basically death sentence + misdiagnosis of HIV had serious personal consequences for a large # of patients.
* U.S. Military was 1 organization that developed a rigorous testing for HIV + used the following procedure for testing recruits.
  + 1 = All applicants were screened w/ an **enzyme-linked immunosorbent assay** (an **ELISA**).
  + If samples tested positive, then 2 more rounds of the same ELISA were performed.
  + If either of those 2 new tests yielded a positive result, then 2 **Western Blot assays** were performed **=** more cumbersome to conduct, but had higher accuracy.
  + Only if both tests were positive did the military determine a recruit to have an HIV infection, based on papers published at the time. For the ELISA, **TP rate**/**sensitivity** of the test was ~93%, + **TN rate/specificity** was ~99%.
  + For the Western blot, **sensitivity** was ~99.9% + **specificity** was ~99.1%.
  + We also know that by the mid 80's, it was estimated that 1.48/1000 adult Americans were HIV positive (**prevalence**).
    1. Quite difficult to track down exact sensitivity + specificity for these tests, or prevalence at the time  these values = approximate, based on what was published at the time.
* Can use Bayes Rule to calculate probability a recruit testing positive in 1st ELISA actually has HIV.
  + Then can consider the sequential testing results.
* Prevalence can be denoted as **probability of having HIV =** .00148
* Sensitivity can be denoted as **Probability of positive** **given HIV =** 0.93
* Specificity can be denoted as **probability of negative given no HIV** = 0.99.
* Prior to any testing, what probability should be assigned for recruit having HIV?
* Given we don't have any additional info about this recruit, our best guess = they are a randomly sampled individual from this population.
* Hence, **prior probability** we assign to this recruit having HIV = simply the **prevalence** of the disease in the population  probability of HIV = 0.00148.
* When a recruit goes through HIV screening, there are 2**competing claims**  recruit has HIV + recruit doesn't have HIV.
* If ELISA yields a positive result, what is the probability this recruit has HIV?
  + Remember, we already decided on the prior probabilities for these 2 competing hypothesis
  + Prior probability of hypothesis that the recruit has HIV = 0.00148 + prior for the competing hypothesis that the recruit does not have HIV = the **complement** of this probability = .99852
* If a recruit *actually* has HIV, there are 2 possible outcomes for the test = positive or negative.
* **Sensitivity** (TP)= probability of a *positive* test result given the person *actually* *has* HIV is 0.93 + the probability of a FN result (probability of testing negative even if person has HIV) = complement of this = 0.07
* Similarly, if a recruit actually does NOT have HIV, there are still 2 possible outcomes ( + and - )
* **Specificity** (TN)= probability of a negative test result given that the person does NOT have HIV = 0.99 + probability of a FP result (probability of testing positive given no HIV) = complement of this = 0.01.
* Remember we're looking for probability a recruit has HIV given they tested positive.
* Using Bayes’ Rule, we can calculate this as probability of HIV + being positive divided by probability of positive.







* By multiplying across the probability tree branches, we can obtain probabilities of HIV given positive, + no HIV given positive.
* To calculate these probabilities, make use of Bayes’ rule again  to obtain probability of HIV given positive, multiply probability of HIV + probability of positive given HIV.
* Probability of HIV + positive = joint
* Overall probability of positive = person can test positive + have HIV or may not have HIV.
* Add these 2 probabilities in the dominator = overall probability of testing positive = marginal
* This yields 0.12 = probability a recruit has HIV given positive on the 1st ELISA = **posterior probability**.
* See probability of having the disease, given a positive test = highly dependent on both the FP + FN rates of the test, as well as the prior probability assumed for the individual.

**Bayes Updating**

* Adopt a Bayesian updating scheme to easily calculate probability of someone actually having  
  HIV given sequential testing results.
* Remember the sequential testing scheme for how early HIV testing worked in the US military.
* If a recruit tests positive, next step = test them again (make a simplifying assumption that sequential tests are independent of each other)
* Since a positive outcome on ELISA doesn't necessarily mean a recruit actually has HIV, they’re retested.
* What is probability of having HIV if the 2nd ELISA also yields a positive result?
* Make probability tree w/ 1st first branch = our priors.
* Remember, this person is *no longer a randomly selected person from the population →* We *know something* about this recruit, that they tested positive on the ELISA once.
* Hence, prior probability assigned to the hypothesis that they have HIV should *change*.
* We **update** our **prior** probability w/ a **posterior** from *the previous test*, calculated as:

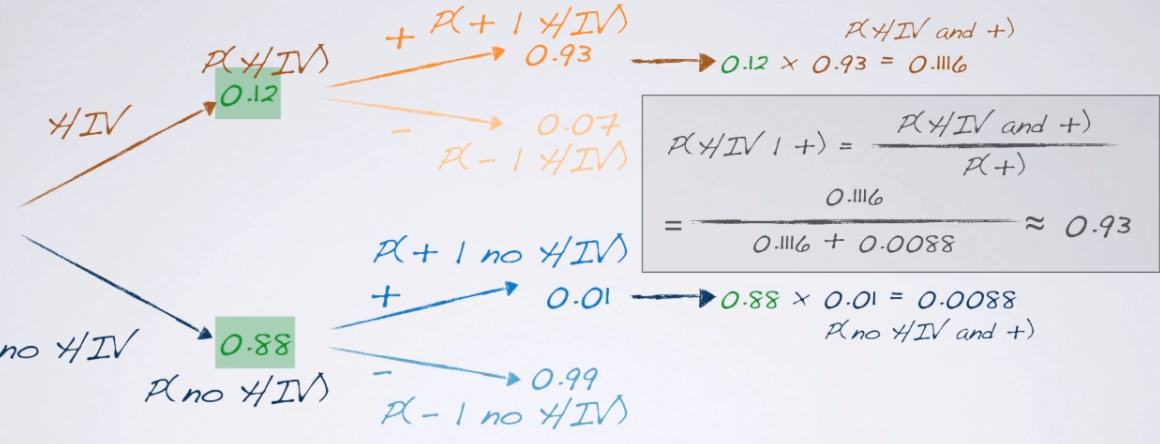
(prob\_HIVGivenPos\_elisa1 <- prob\_posAndHIV\_elisa / prob\_pos\_elisa1)

**[1] 0.1211449**

* The prior probability for the competing hypothesis that a recruit does NOT have HIV also gets updated as the *complement* of this probability.
* Nothing should change in the 2nd branch b/c we're testing the SAME TEST w/ SAME sensitivity + SAME specificity.

(prob\_HIVGivenPos\_elis2 <- prob\_posAndHIV\_elisa2 / prob\_pos\_elisa2)

**[1] 0.9276384**



* Updated posterior probability comes out to 0.93 = A recruit who tested positive on ELISA twice has a 93% chance of having HIV.
* The calculated posterior probability of having HIV after 3 consecutive positive ELISA’s should be more than enough for any **individual** diagnostic decision.
* However, it's important to realize the military was testing hundreds of thousands of recruits, hence their need for *additional accuracy* provided by **western blots**.
* So if a recruit tested positive on 3 consecutive ELISA tests, the next step = test them w/ a western blot, a different test w/ a different sensitivity + specificity.
* Hence in addition to updating the prior, also need to update probabilities in the 2nd branches.
* This updating scheme = example of a general property of Bayesian models.

# **Bayesian vs. Frequentist Definitions of Probability**

* Compare Bayesian approaches to **frequentist approaches**.
* Consider these 3 statements
* Probability of flipping a coin + getting heads = ½
* Probability of rolling snake eyes (two 1s on 2 dice) = 1/6 \* 1/6 = 1/36
* Probability of Apple's stock price going up today = 0.75.
* What *exactly* do these statements *mean*?
* How you interpret these statements depends on your *definition of probability*.
* 1 definition of probability of an event = its **relative frequency** in a large # of trials.
* If you can repeat flipping a coin indefinitely + count how many heads you get + divide by # of flips, the value you obtain should be 0.5.
* In other words, probability of event E = the proportion of times event E occurs in n trials when n goes to infinity.
* This is the **frequentist definition of probability**,
* Suppose now you're indifferent between winning a $1 if event E occurs where event E = drawing a blue chip from a box w/ 1,000 x p blue chips + 1,000 x (1-p) white chips.
  + Or, probability of event E, **P(E) =** probability of drawing a blue chip from this box --> **P(E) = p**
* This definition of probability is based on your **degree of belief =** the **Bayesian definition**.
* Earlier courses, we talked about **frequentists methods of inference**, such as CI
* When defining the **confidence level**, we were *very careful* to describe it as the proportion of random samples of size n from the same population that produced CI's that contain the true population parameter.
* We emphasized that an interpretation of the confidence level as "probability that a given interval containing the true parameter" is INCORRECT.
  + Ex: 2015 Pew Research poll on 1500 adults --> got a CI that said "We're 95% confident that 60-64% of Americans think the federal government does not do enough for middle class people"
  + In this statement, "95%" means 95% of random samples of 1,500 adult Americans will produce CI's for the proportion of Americans who think the federal government does not do enough for a middle-class people that will contain the true proportion
* Some common misconceptions about the confidence level:
  + Interpret it value as "there's a 95% chance that this CI includes the true population proportion.
  + Or "the true population proportion is in this interval 95% of the time"
  + The frequentist definition of probability allows us to define a probability for the CI procedure *but not for specific fixed sample.*
  + In the case of a specific fixed sample, when the data do NOT change, we will either *always* capture the true parameter or *never* capture it.
  + In other words, for given CI, the true parameter is either in it or not.
  + This is the same as saying "the probability a given CI captures the true parameter is either 0 or 1"
  + The only problem here is that we *can't know whether the probability that this given interval captures the true parameter is 0 or 1*.
  + The Bayesian definition is a bit more flexible, since it's a measure of belief which allows us to describe the unknown true parameter NOT as a *fixed value,* but w/ a *probability distribution*.
  + This lets us construct something like a CI, except we will also be able to make **probabilistic statements** about the parameter falling w/in that range.
  + Ex: Could say something like "Posterior distribution yields a 95% **credible interval** of 60-64% for proportion of Americans who think the federal gov. does not do enough for middle class people.

**Inference for a Proportion: Frequentist Approach**

* Study addressed the question of whether a controversial drug RU-486 could be an effective morning-after contraceptive.
* Study participants = 40 women who came to a health clinic asking for emergency contraception.
* Investigators randomly assigned 20 women to receive RU-486 + other 20 to receive standard therapy, consisting of high doses of the sex hormone estrogen + a synthetic version of progesterone.
* Of the women assigned to RU-486 (treatment group) + 4 became pregnant.
* Of the women who received the standard therapy (control), 16 became pregnant.
* How strongly do these data indicate the treatment is more effective than the control?
* Framework.
  + To simplify matters, turn this problem of comparing 2 proportions to a *1-proportion problem*.
  + Consider the 20 total pregnancies + ask *how likely is it 4 come from the treatment group?*
  + If the treatment + control are equally effective + the sample sizes for the 2 groups are the same, then the probability a pregnancy came from the treatment group is simply 0.5.
* In the frequentist approach, we 1st need to *set our hypotheses*.
* But before that, define the parameter of interest to use in these hypotheses.
* Let p = probability a given pregnancy comes from the treatment group.
* Then H0: p = 0.5 🡺 no difference between treatment + control groups + the pregnancy is equally likely to come from either group.
* H1: p < 0.5 🡺 treatment is more effective + a pregnancy is less likely to come from treatment group.
* To make a decision w/in the frequentist paradigm, we need a **p-value** = probability of the observed or more extreme outcome (in direction of the alternative) given the null is true.
* Outcome in this experiment = 4 “successes” (**k**) in 20 trials (**n**)
* The null states probability of success = .5, so we calculate the p-value as probability of obtaining 4 or fewer successes in 20 trials where probability of success is 0.5.
* This probability can be calculated exactly w/ a binomial distribution.

> k <- 4

> n <- 20

> sum(dbinom(0:k,n,.5))

[1] 0.005908966

* Remember: **The # of successes in a fixed # of independent trials for a categorical random variable with 2 levels that can be defined as a success or a failure follows a binomial distribution, w/ 2 parameters, n + p**.
* In this case n = 20 + p = 0.5, but we're looking for 4 or fewer successes, defined as probability k *is at most* 4.
* We calculated this probability = 0.0059 = chances of observing 4 or fewer pregnancies in the treatment group, given pregnancy was equally likely in the 2 groups (null is true)
* With such a small probability, we’d reject the null + conclude the data provide convincing evidence for the treatment being more effective than the control.

**Inference for a Proportion: Bayesian Approach**

* Now to answer the question using a Bayesian approach 🡪 how likely is it that 4 pregnancies occur in the treatment group?
* We had decided that if the treatment + control are equally effective, + the sample sizes for the 2 groups are the same, then the probability the pregnancy comes from the treatment group is 0.5.
* Within the Bayesian framework, we again start by setting our hypotheses + think of these as the models that the data come from.
* Begin by delineating each of the models we consider **plausible**.
* We know p (probability a pregnancy comes from the treatment group) can take on any value between 0 + 1
* Start slow + instead of considering a *continuous* parameter space for p, assume it’s plausible that chances a pregnancy comes from the treatment group is 10% or 20% or 30% or 40%, up to 90%.
* Hence, we're considering 9 models, compared to just 1 model for the classical frequentist paradigm.
* For p = 20%, this means that, *given a pregnancy occurs*, there is a 2:8, or 1:4, chance it will occur in the treatment group.
* Next, we need to specify the **prior probabilities** to assign to hypotheses.
* The prior probabilities should reflect our **state of belief prior to the current experiment +** should incorporate info learned from all relevant research up to the current point in time.
  + However, *do not incorporate info from the current experiment*.



* We placed gave p = 0.5 a prior probability of 52% + equally divided the remaining probability among the other models.
* This equal distribution implies that the benefit of the treatment is symmetric/is equally likely to be better or worse than the standard treatment.
* The 52% prior at p = 0.5 implies we believe there's a 52% chance that there is no difference between the treatments.
* 1 natural question to ask at this point is *how did you come up with those priors*?
* For now, stick w/ the chosen priors + work through the mechanics of calculating the posterior probabilities + making a decision.
* Now we're ready to calculate the probability of the observed data, given each model
* This probability is called the **likelihood** = probability of the data, given the model
* Here, this is probability that k = 4, given that n = 20 + various values of p decided to consider as plausible models (10% through 90%)



* As w/ frequentists, express the probability of a given # of successes in a given # of independent trials **via a binomial distribution.**
* We consider a sequence of probabilities of success from 10%-90%, increasing by 10%, + assign a 52% prior probability to p = 0.5, + 6% probabilities to all other models.
  + We *won't actually use these prior probabilities in the calculation of the likelihood*, but they’ll become relevant for the calculation of the posterior
* Finally, calculate the likelihood as a binomial w/ 4 successes + 20 trials, when p = the variety of values we're considering.

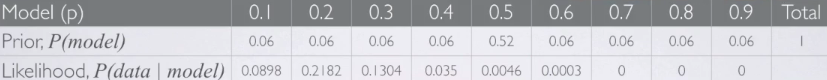
> p\_values <- seq(.1,.9,.1)

> priors <- c(rep(.06,4), .52, rep(.6,4))

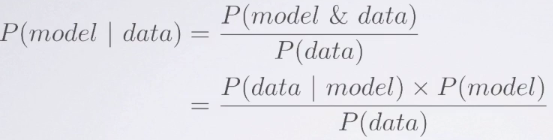
> (likelihood <- dbinom(4, 20, prob = p\_values))

[1] 8.977883e-02 2.181994e-01 1.304210e-01 3.499079e-02 4.620552e-03 2.696862e-04

[7] 5.007558e-06 1.300570e-08 3.178804e-13



* The # of successes + trials are the same for each of these likelihoods, but each likelihood uses a different probability of success, based on which model it’s based on.
* Once the models are delineated, priors are expressed + the data are collected, we can use a re-expression of Bayes' rule to calculate posterior probability = probability of the model given the data.



* Probability of model, given data = probability of model + data divided by probability of data (overall, considering all possible models)
* We further write the numerator as probability of data given the model multiplied by prior probability of that model.
* To get posteriors = multiply vector of priors, **P(model)**, defined earlier w/ likelihoods, **P(data|model)**
* **P(data)** = sum of the probabilities for the various models

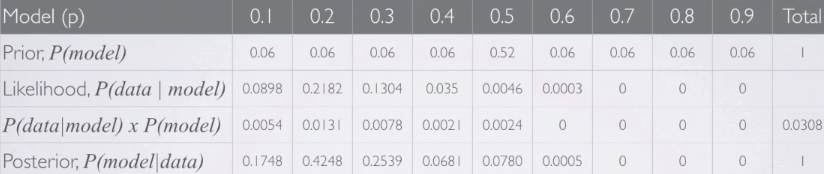
> (posteriors <- (priors \* likelihood) / sum((priors \* likelihood)))

[1] 1.739287e-01 4.227181e-01 2.526648e-01 6.778772e-02 7.757883e-02 5.224635e-03 9.701152e-05 2.519596e-07 6.158304e-12

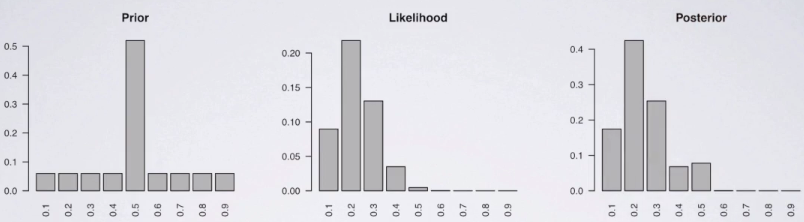
* This mimics the calculation based on probability trees seen before, where this denominator sums up all possible probabilities where the data might be coming from.
* Also check to make sure the posterior probabilities add up to 1,

> sum(posteriors)

[1] 1



* See that the posterior probability is highest at p = 0.2, so this model is the most likely model, based on the observed data.
* Even though we assigned a low prior to this model (.06), **the incorporation of the data gave this model a high probability.**
* This shouldn't be surprising, since 4 successes in 20 trials is ~20%.
* So the calculation of the posterior **incorporated prior info + likelihood** **of the data observed**
  + **The concept of data “at least as extreme as observed” placed no part in the Bayesian paradigm.**
* Finally, note the probability p = 0.5 dropped to about 7.8% in the posterior.
* This demonstrates how we **update beliefs based on observed data**.



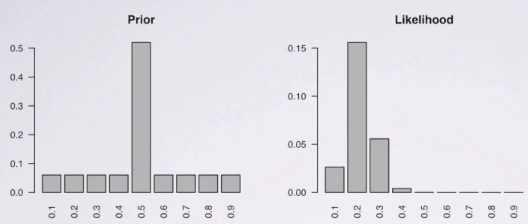
* The Bayesian paradigm, unlike the frequentist approach, also allows us to make **direct probability statements** about our models.
* For example, we can calculate probability RU-486 (treatment) is more effective than the control as the sum of the posteriors of the models where p < 0.5.



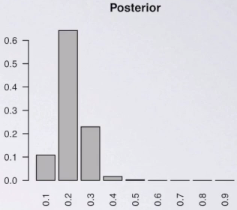
* There is a 92.16% chance that the treatment is more effective than the control.

**Effect of Sample Size on the Posterior**

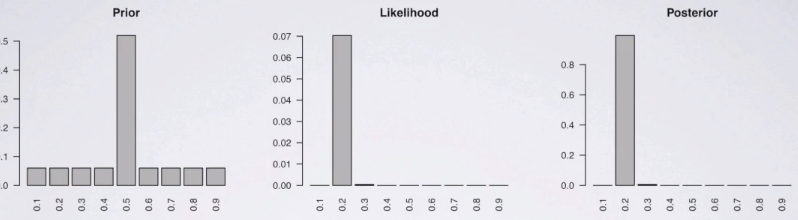
* So we’ve obtained a posterior distribution w/ the highest posterior probability @ p = 0.20.
* While there's a peak at this value in the distribution, there's still some uncertainty in the posterior, as other models also have some probability mass in the posterior distribution.
* Look at what the posterior distribution would look like if we had more data, suppose n = 40 + k = 8
  + Note we're still maintaining the 20% ratio between sample size + number of successes.
* Start w/ the same prior distribution 🡺 Calculate likelihood of the data which



* Likelihood is again centered at 0.20, but is less variable than the original likelihood w/ the smaller n
* Finally put these two together to obtain the posterior distribution.



* The posterior also has a peak at p = 0.20, but the peak is taller = there is more probability mass for that model + less on the others
* Still keeping the 20% ratio between sample size + number of successes, consider n = 200 + k = 40 w/ the same prior



* Likelihood is again centered at 20% + almost all of probability mass in the posterior is at p = 0.20.
* The other models do not have 0 probability mass, but they're posterior probabilities are very close to 0.
* So as you can see, as more data are collected, likelihood ends up dominating prior.
* This is why, while a good prior helps, a **bad prior can be overcome w/ a large sample**.
  + *This only works if we don't place a 0 probability mass on any of the models in the prior*