***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**

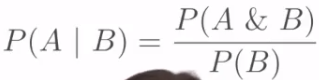
* 2015 Gallup Poll asked adult Americans whether they have personally ever used an online dating site, such as Match.com, eHarmony, or OKCupid.
* See distribution of responses by age group in a contingency table.



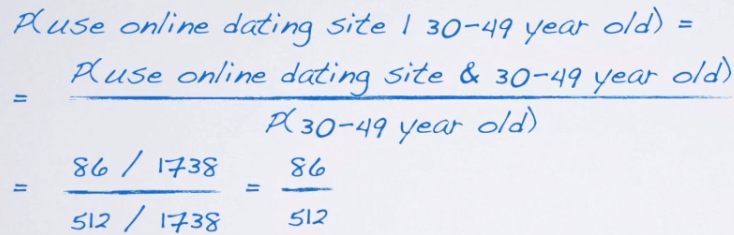
* Using these data, calculate what % of 30-49-year-olds use online dating sites  86/512 = ~17%.
* Formally, this = probability of using online dating sites **given 30-49 years old**.

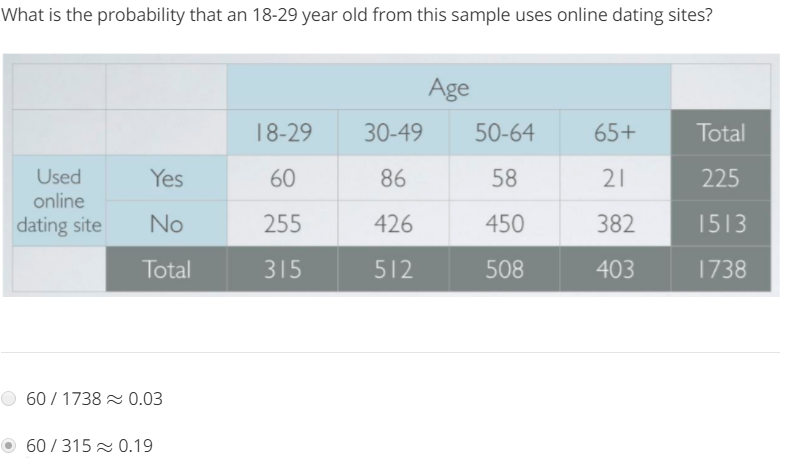


* Calculated this probability simply as a ratio of 2 frequencies, but we can formalize things a bit more
* Let event A = using an online dating site, event B = being 30-49 years old.



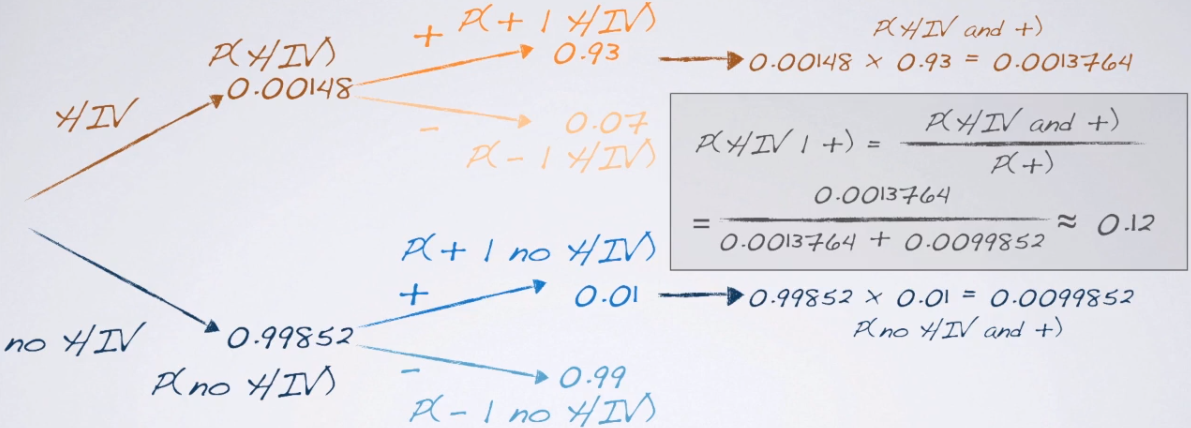
* Numerator frequency = # of times events A + B happened at the same time
* Denominator = # of times event B happened.
* Probability of A given B = probability of A + B divided by probability of B = **Bayes' rule**
* Thomas Bayes (1702-1761) = mathematician who established a mathematical basis for **probability 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%.

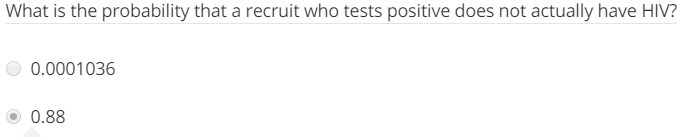


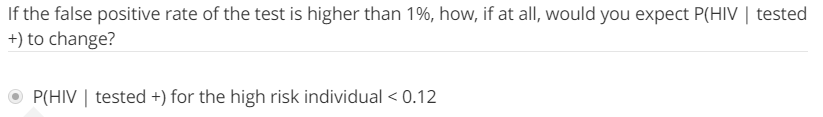


**Bayes' Rule and Diagnostic Testing**

* 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:

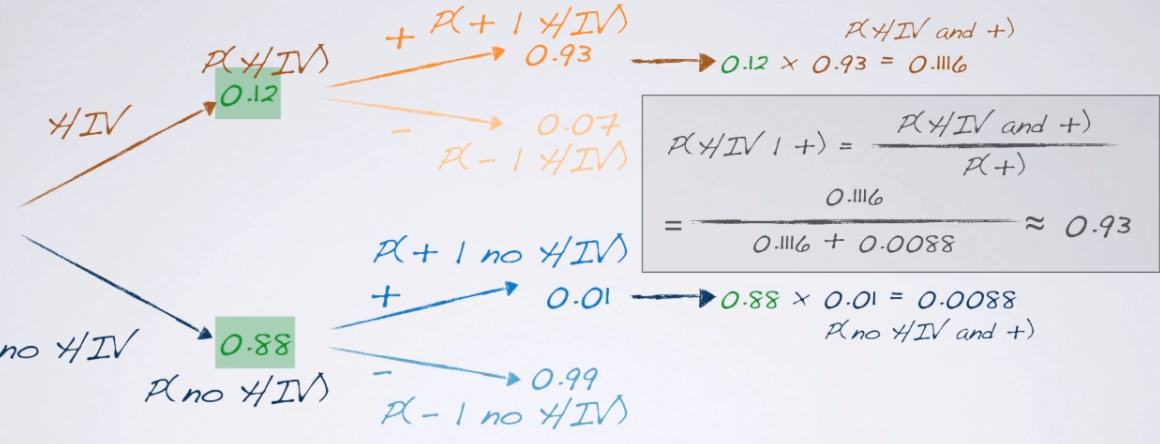
(prob\_HIVGivenPos\_elisa1 <- prob\_posAndHIV\_elisa / prob\_pos\_elisa1)

**[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 the proportion of Americans who think the federal government does not do enough for middle class people.