***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**

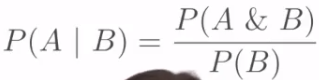
* 2015 Gallup Poll asked adult Americans whether they have personally ever used an online dating site, such as Match.com, eHarmony, or OKCupid.
* See distribution of responses by age group in a contingency table.



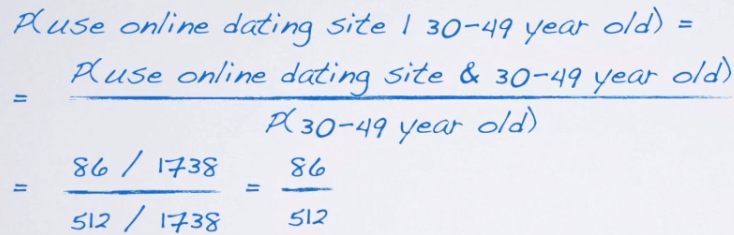
* Using these data, calculate what % of 30-49-year-olds use online dating sites 🡺 86/512 = ~17%.
* Formally, this = probability of using online dating sites **given 30-49 years old**.

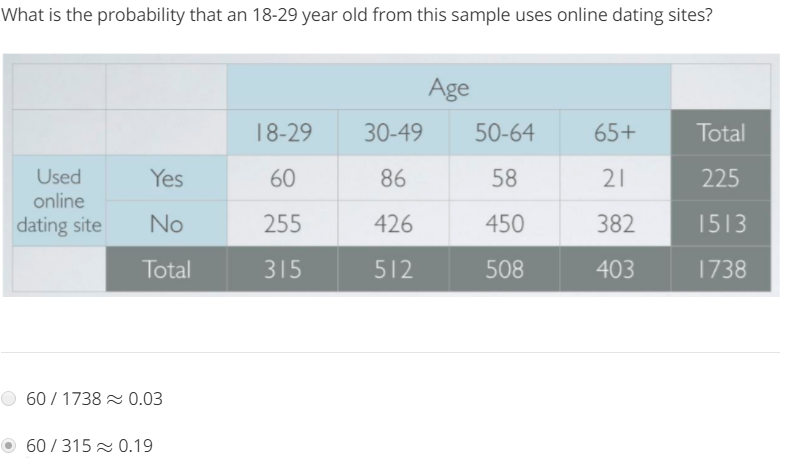


* Calculated this probability simply as a ratio of 2 frequencies, but we can formalize things a bit more
* Let event A = using an online dating site, event B = being 30-49 years old.



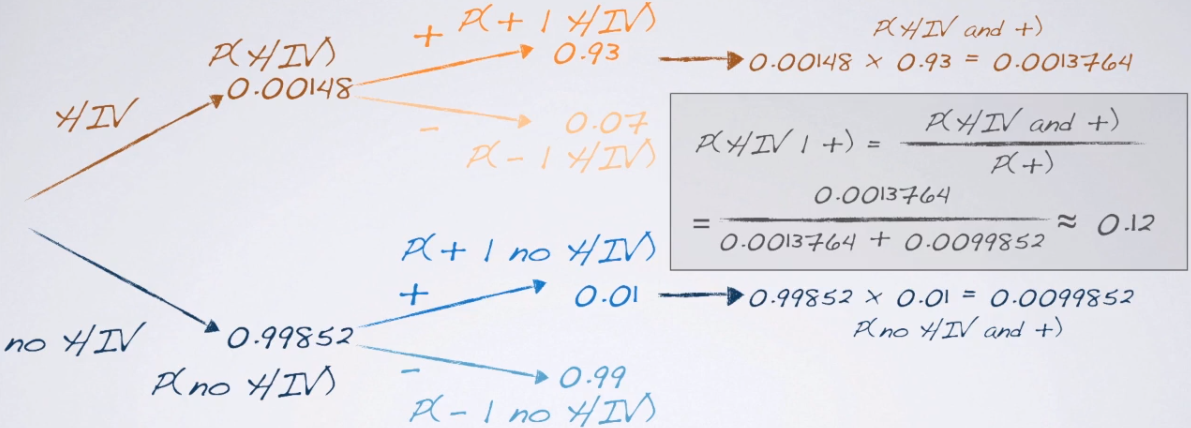
* Numerator frequency = # of times events A + B happened at the same time
* Denominator = # of times event B happened.
* Probability of A given B = probability of A + B divided by probability of B = **Bayes' rule**
* Thomas Bayes (1702-1761) = mathematician who established a mathematical basis for **probability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