Starting off

Emmitt is 6'6" and very athletic. Emmitt graduated from college in three years and then went on to start his career. Based on the information you have, which do you think is more likely Emmitt plays in the NBA or Emmitt is an accountant?

Bayesian Statistics

Data Science Immersive

FLATIRON SCHOOL

Frequentist vs. Bayesian

- What percentage of NBA players are above 6’6”?
- What percentage of accountants are about 6’6”?
- What percentage of the population plays in the NBA?
- What percentage of the population are accountants?
- While we have some evidence that makes it more likely that the person plays in the NBA, we still have to start with our prior knowledge that very few people overall play in the NBA.

Frequentist vs. Bayesian

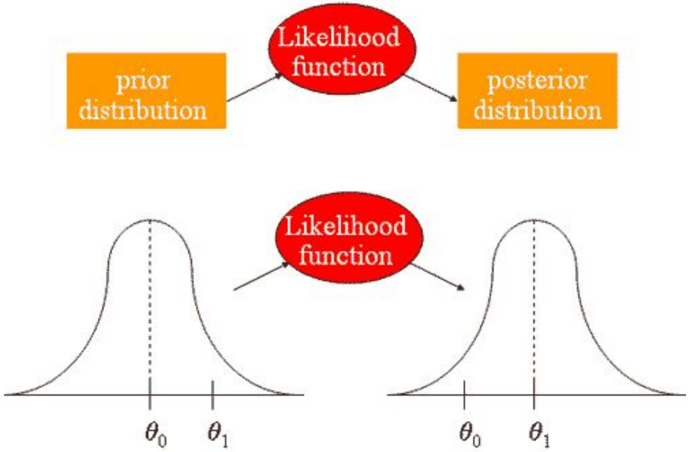
- Frequentist statisticians rely on the imaginary sampling of an infinite population and derive a probability value that summarizes the result of the experiment
  - Inference made by a frequentist statistician only depends on the frequency of events, or samples observed
- Bayesian statisticians not only rely on actual evidence observed, but also on beliefs

Bayesian Statistics

Bayesian statistics is a particular approach to applying probability to statistical problems.

We may have a *prior* belief about an event, but our beliefs are likely to change when new evidence is brought to light. Bayesian statistics gives us a solid mathematical means of incorporating our prior beliefs, and evidence, to produce new *posterior* beliefs.

In particular Bayesian inference interprets *probability* as a measure of *believability* or *confidence* that an *individual* may possess about the occurrence of a particular event.



Frequentist vs. Bayesian

Situation: You flip a coin 10 times and it comes up heads 7 out of the 10 times. What is the probability of seeing a heads come on your next flip?

Frequentist Interpretation	Bayesian Interpretation
The probability of seeing a head when the coin is flipped is the *long-run relative frequency* of seeing a head when repeated flips of the coin are carried out. That is, as we carry out more coin flips the number of heads obtained as a proportion of the total flips tends to the "true" or "physical" probability of the coin coming up as heads. In particular the individual running the experiment *does not* incorporate their own beliefs about the fairness of other coins.	Prior to any flips of the coin an *individual may believe* that the coin is fair. After a few flips the coin continually comes up heads. Thus the *prior* belief about fairness of the coin is modified to account for the fact that three heads have come up in a row and thus the coin might not be fair. After 500 flips, with 400 heads, the individual believes that the coin is very unlikely to be fair. The *posterior* belief is heavily modified from the *prior* belief of a fair coin.

Frequentist vs. Bayesian

Situation: Election of a candidate

Frequentist Interpretation	Bayesian Interpretation
The candidate only ever stands once *for this particular election* and so we cannot perform "repeated trials". In a frequentist setting we construct "virtual" trials of the election process. The probability of the candidate winning is defined as the relative frequency of the candidate winning in the "virtual" trials as a fraction of all trials.	An *individual* has a *prior* belief of a candidate's chances of winning an election and their confidence can be quantified as a probability. However another individual could also have a separate differing prior belief about the same candidate's chances. As new data arrives, both beliefs are (rationally) updated by the Bayesian procedure.

Frequentist vs. Bayesian

***Frequentist*:** "If a very large number of samples, each with the same sample size as the original sample, were taken from the same population as the original sample, and a 95% confidence interval constructed for each sample, then 95% of those confidence intervals would contain the true value of θ." This is an extremely awkward and dissatisfying definition but technically represents the frequentist's approach.

***Bayesian*:** "The 95% confidence interval defines a region that covers 95% of the possible values of θ." This is much more simple and straightforward. (As a matter of fact, most people when they first take a statistics course believe that this is the definition of a confidence interval.)

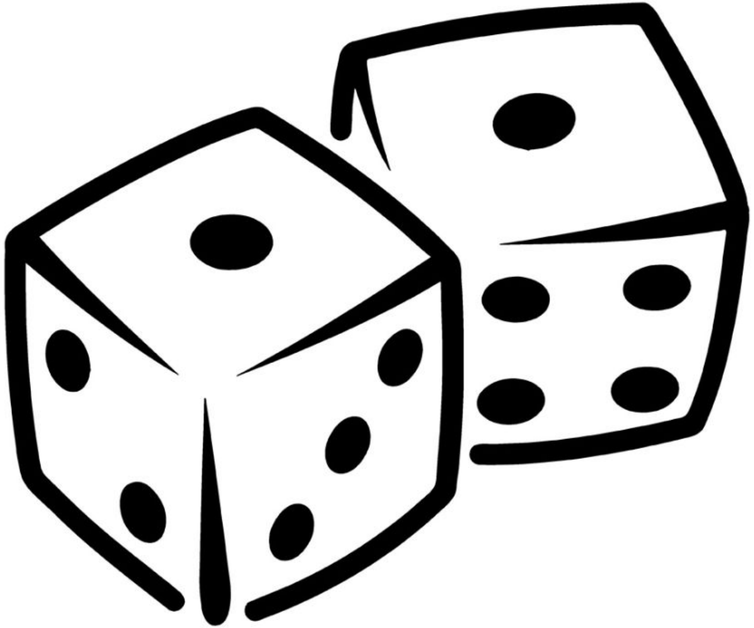
In order to carry out Bayesian inference, we need to utilize a famous theorem in probability known as Bayes' rule and *interpret it in the correct fashion*.

Review

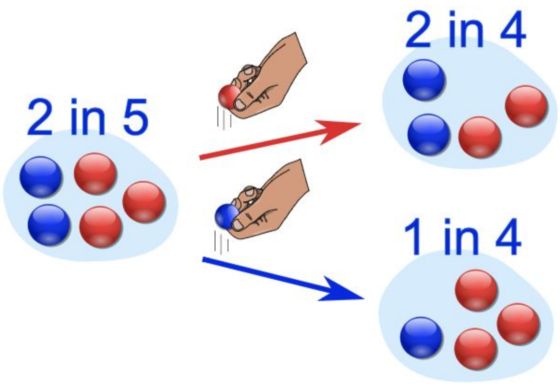
- Event
    - An **event** is the outcome of a random experiment
- Sample space
    - A **sample space** is a collection of every single possible outcome in a trial
- Independent probability is calculated by event divided by all the possible events in the sample space
    - The occurrence of one event does not affect, or dependent on the outcome of another

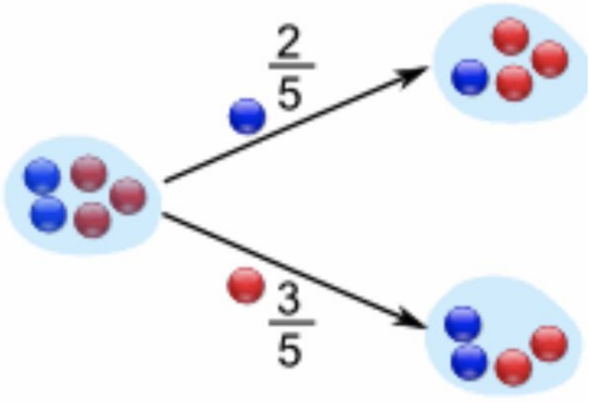
Review

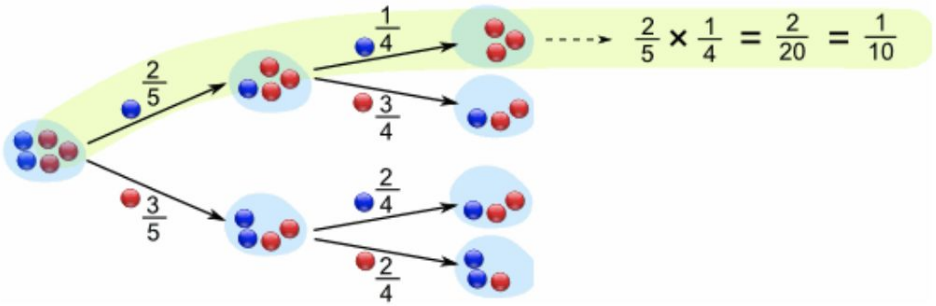
- Independent probability of an event is calculated as P(A) divided by **all** possible events.
- Some statistical distributions make such assumptions about events
  - Poisson distribution
  - Binomial distribution



Conditional Probability

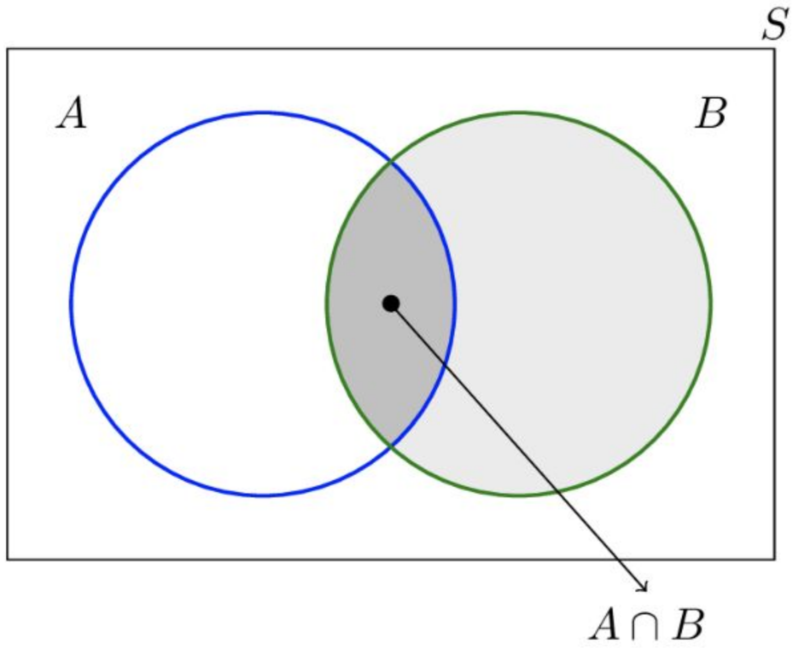






Conditional Probability

- Conditional Probability emerges in the examination of experiments where a result of a trial may influence the results of the upcoming trials. For example:
  - The probability of drawing an Ace given already drew an Ace
  - The probability of being in a good mood given the weather is nice



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example 1

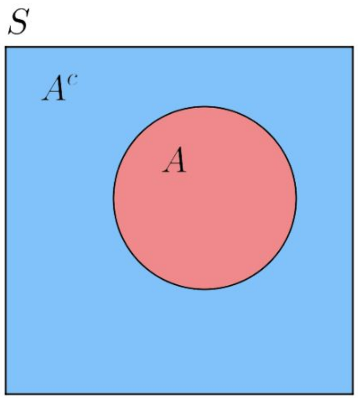
You are going to visit your distant cousins who recently had two children. You are told that at least one of them is a girl. What is the probability of both of them being girls?

Example 2

You are going to visit your distant cousins who recently had two children. You are told that the older one is a girl. What is the probability of both of them being girls?

Theorems of Conditional Probability

1. P(A’) + P(A) = 1



Sample Space S, event A, and complement A^c

2. The Product Rule

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

The product rule is useful when the conditional prob is easy to compute but the intersection is not

Bayes' Theorem

- Bayes' Theorem underlies the foundation of Bayesian Inference, an incredibly powerful way in which statistics are probability are computed
- It is derived from conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Bayes' Theorem

P(A) is called the **prior**; this is the probability of our hypothesis without any additional prior information. It could also be a belief we have prior to seeing the data.

P(B) is called the **marginal likelihood**; this is the total probability of observing the evidence. In many applications of Bayes Rule, this usually serves as normalization constant, which we will explain in further detail.

P(B|A) is called the **likelihood**; this is the probability of observing the new evidence, given our initial hypothesis.

P(A|B) is called the **posterior**; this is what we are trying to estimate.

Bayes' Theorem

Bayes' Theorem

Suppose we have events $A_1, \dots, A_k$ and event B. If $A_1, \dots, A_k$ are k mutually exclusive events, then...

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)} = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k)}$$

Applying this to just two events A and B we have...

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')}$$

Bayes Theorem Example

In a jury trial, suppose the probability the defendant is convicted, given guilt, is 0.95, and the probability the defendant is acquitted, given innocence, is 0.95. Suppose that 90% of all defendants truly are guilty. Find the probability the defendant was actually innocent given the defendant is convicted.

Let Guilty = G

Innocent = I

Acquitted = A

Convicted = C

$P(G) = 0.9$ so $P(I) = 0.1$

$P(C|G) = 0.95$ so $P(A|G) = 0.05$

$P(A|I) = 0.95$ so $P(C|I) = 0.05$

Need to find: $P(I|C)$

Bayes Theorem Example

In a jury trial, suppose the probability the defendant is convicted, given guilt, is 0.95, and the probability the defendant is acquitted, given innocence, is 0.95. Suppose that 90% of all defendants truly are guilty. Find the probability the defendant was actually innocent given the defendant is convicted.

$$\begin{aligned} P(I \text{ and } C) &= P(C|I) * P(I) \\ &= 0.05 * 0.1 \\ &= 0.005 \end{aligned}$$

$$\begin{aligned} P(C) &= P(G \text{ and } C) + P(I \text{ and } C) \\ &= (0.95) * (0.9) + (0.05) * (0.1) \\ &= 0.855 + 0.005 \\ &= 0.86 \end{aligned}$$

Bayes Theorem Example

In a jury trial, suppose the probability the defendant is convicted, given guilt, is 0.95, and the probability the defendant is acquitted, given innocence, is 0.95. Suppose that 90% of all defendants truly are guilty. Find the probability the defendant was actually innocent given the defendant is convicted.

$$
\begin{aligned}
P(I|C) &= \frac{P(I \text{ and } C)}{P(C)} \\
&= \frac{0.005}{0.86} \\
&= 0.006
\end{aligned}
$$

Bayes Theorem Practice

You are planning a picnic today, but the morning is cloudy.

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is an usually dry year with only rain on 10% of days.

What is the chance of rain during the day?

More practice

https://www.mathsisfun.com/data/bayes-theorem.html

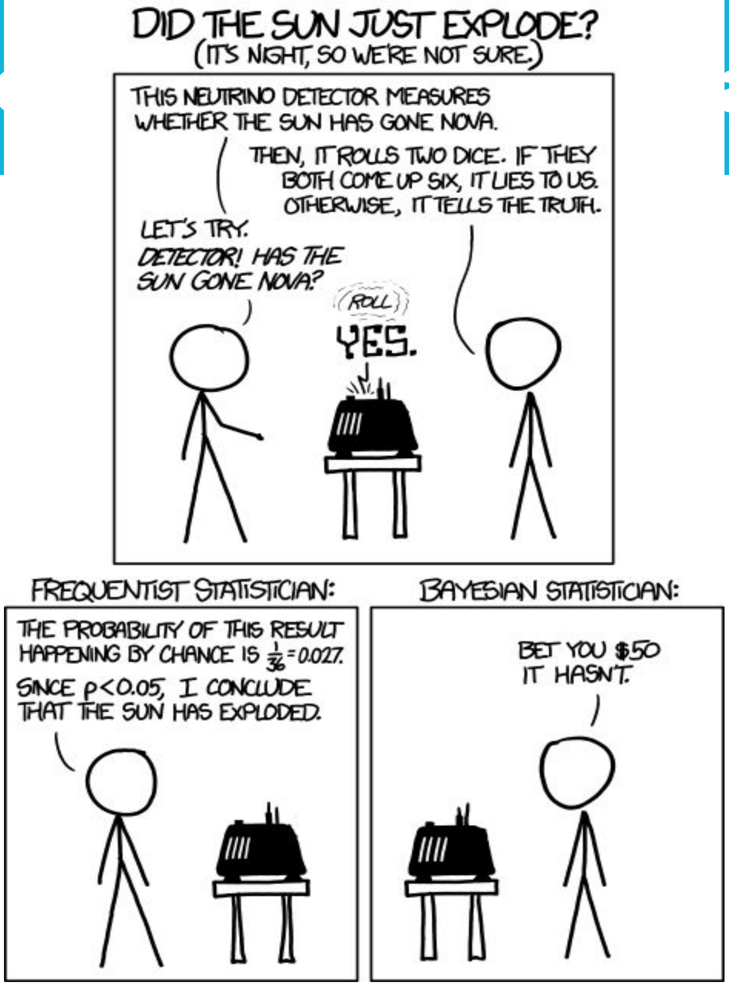
Bayes' Theorem

Why Bayes'?

- Bayes' theorem allows us to accommodate degrees of belief into the equation, which accounts for uncertainty
- Has practical implication in natural language processing--i.g. Naive bayes for classification or topic modeling, such as Latent Dirichlet Allocation
- Stochastic methods for parameter estimation, i.g. Markov Chain Monte Carlo
- Network analysis or graph theory
- And many more applications in philosophy, cognitive sciences, even legal systems

Great Tutorial from 3 Blue 1 Brown.
https://www.youtube.com/watch?v=HZGCoVF3YvM

Fre an


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.