DataSci 520

lesson 7

Bayesian statistics

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

today's agenda

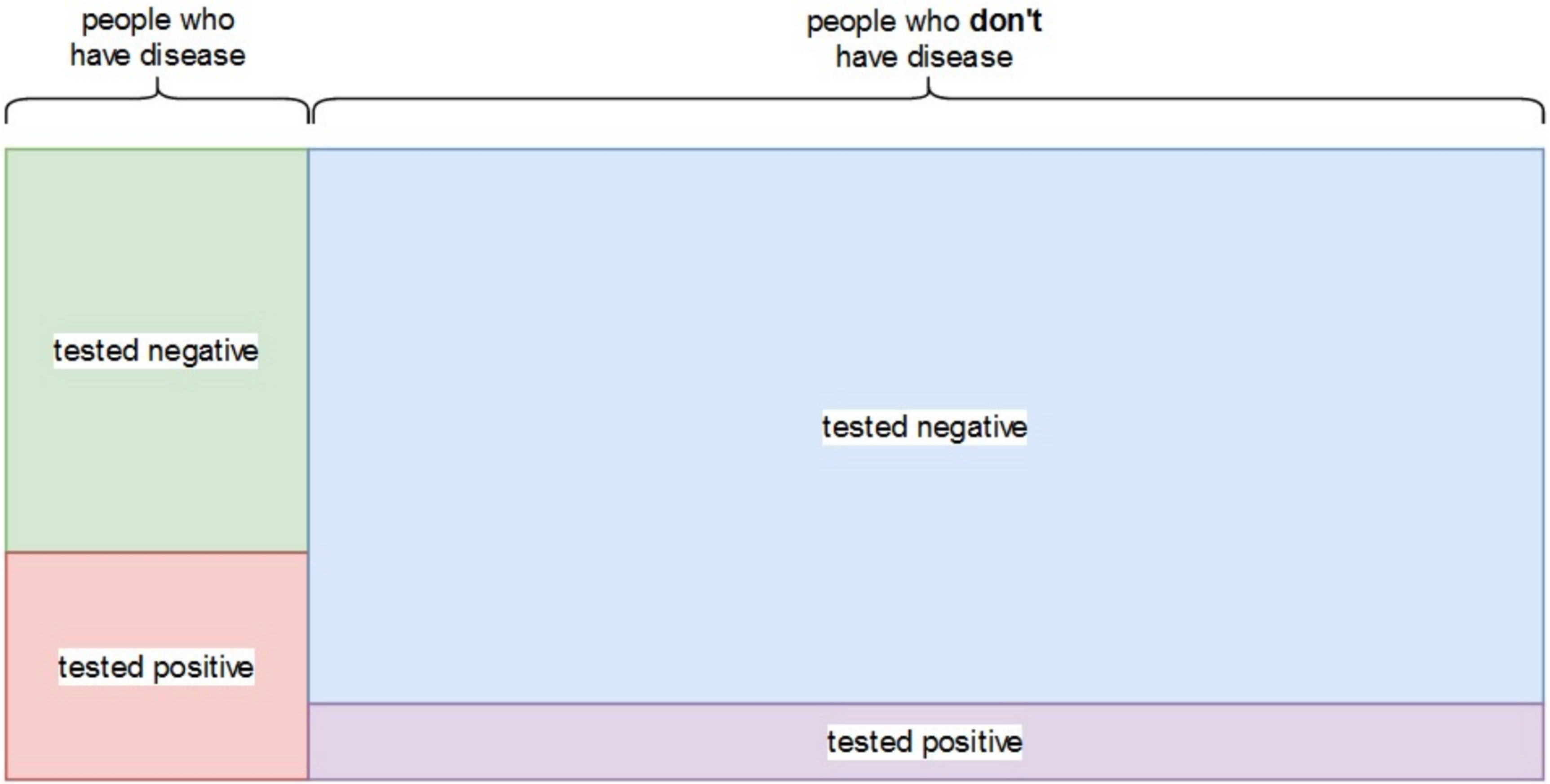
- Bayes' rule and its implication
- Bayesian hypothesis testing
- prior, likelihood and posterior
- Frequentist vs Bayesian paradigms
- the billiard game

Bayes' theorem motivation

Sometimes $P(A|B)$ is not easy to get but $P(B|A)$ is (or vice versa), for example

- if A is testing positive and B is having a disease, then $P(A|B)$ depends only the the test's accuracy, but $P(B|A)$ depends on the test *and* on the prevalence of the disease in the population
- if A is the event that I toss a coint 4 times and get 3 heads and B is the event that my coin is un-biased, then $P(A|B)$ can be looked up from the binomial distribution with $n = 4$ and $p = 0.5$, but $P(B|A)$ is not so easy

Bayes' theorem gives us a formula for calculating $P(A|B)$ in terms of the $P(B|A)$.



P(A|B) = 🟥 / (🟥 + 🟩)

P(B|A) = 🟥 / (🟥 + 🟪)

Which of the above two probabilities is more important if you are a patient?

Bayes' theorem

Bayes' theorem is derived from the conditional and marginal probabilities:

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

and from conditional probability we know that

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

which we can use to rewrite the above equations as

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

and by symmetry $P(A \cap B) = P(B \cap A)$ we can also switch A and B

two ways to simplify

- $P(B)$ is something we usually don't know, so we just guess it and most of the time a good guess is enough (however critics of of Bayesian analysis usually point to this as a weak point)
- $P(A)$ is obtained by summing (or integrating) over the sample space of B:

  $$P(A) = P(A|B)P(B) + P(A|B')P(B')$$

  where B' is the **complement** of B, however this cumbersome calculation is often skipped in practice because it is not needed, which is why Bayes' formula is also sometimes written as

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

  with the denominator gone, and the equality replaced with $\propto$ (proportional to)

prior, likelihood and posterior distribution

Let's rewrite Bayes' formula using D and H this time, where

- H stands for some hypothesis, i.e. some distributional assumption about the parameter(s) (we ne longer assume parameters to be fixed!)
- D stands for data, i.e. **evidence** for or against the hypothesis

$$\overbrace{P(H|D)}^{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}}\overbrace{P(H)}^{\text{prior}}}{\underset{\substack{\uparrow \\ \text{not very important to get this right} \\ \text{(it is just a normalization factor)}}}{P(D)}}$$

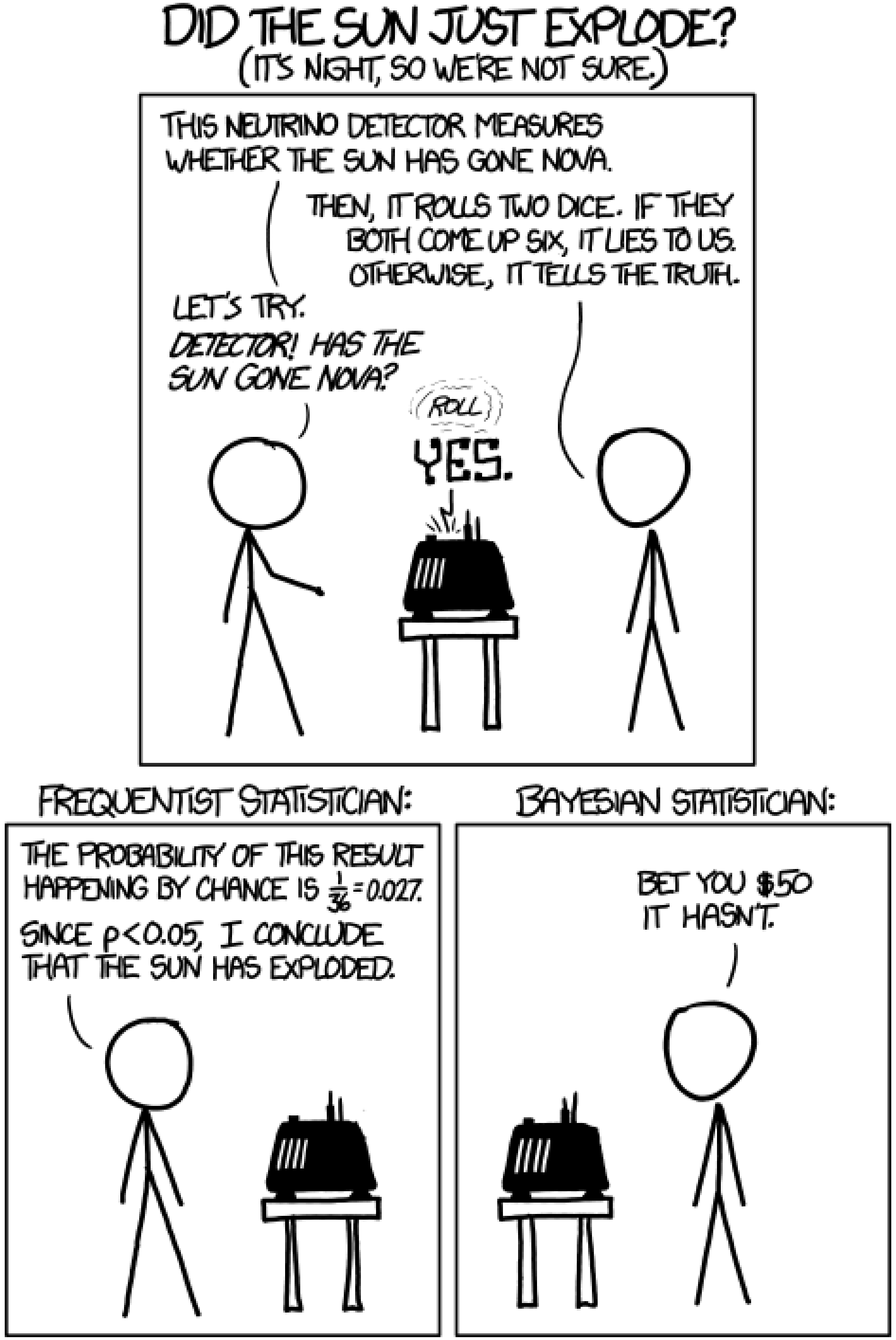
Frequentist vs Bayesian paradigm

There is a rift among statisticians about how to interpret a probability:

- **Frequentist:** in terms of relative frequencies of repeated events (sampling data) given fixed parameters, for example
    - in classical hypothesis testing (previous lesson), we let $H = H_0$ be the null hypothesis and $P(D|H_0)$ is the likelihood (under H_0), if the p-value is less than α, we reject H_0, otherwise we fail to reject H_0
- **Bayesian:** in terms of uncertainty around parameter values *prior to and after* observing data, for example:
    - in Bayesian hypothesis testing, once we compute the posterior $P(H|D)$ and we can use it to test any hypothesis, not just H_0, so there is **no null and alternative hypotheses**, and **no p-value**

Bayesian View

- Most of the statistics we have been doing rely on assumed parameters and limiting distributions. This is called “Frequentist Statistics”.
- The main different between Bayesian and Frequentist statistics is that a Bayesian view of the world includes updating/changing our beliefs when we observe data along with taking into account prior beliefs.
- Example: If we’ve lost our keys, we either:
  - (1) Search our house from top to bottom.
  - (2) Search our house starting at the areas we have previously lost our keys before (laundry basket, desk, pockets, …), then we more into more and more less likely places.
- Using a specific way to solve some problems does not require you to sign up for a lifetime of using that exact way. In fact, the common belief is that some problems are better handled by Frequentist methods, and some with Bayesian methods.
- What is the “controversy”?
  - Some Bayesian methods use “priors” to quantify what we know about parameters.
  - Frequentists do not quantify anything about the parameters, using p-values and confidence intervals to express unknowns.


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 $\frac{1}{36}=0.027$.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

An overused example, but for good reason.

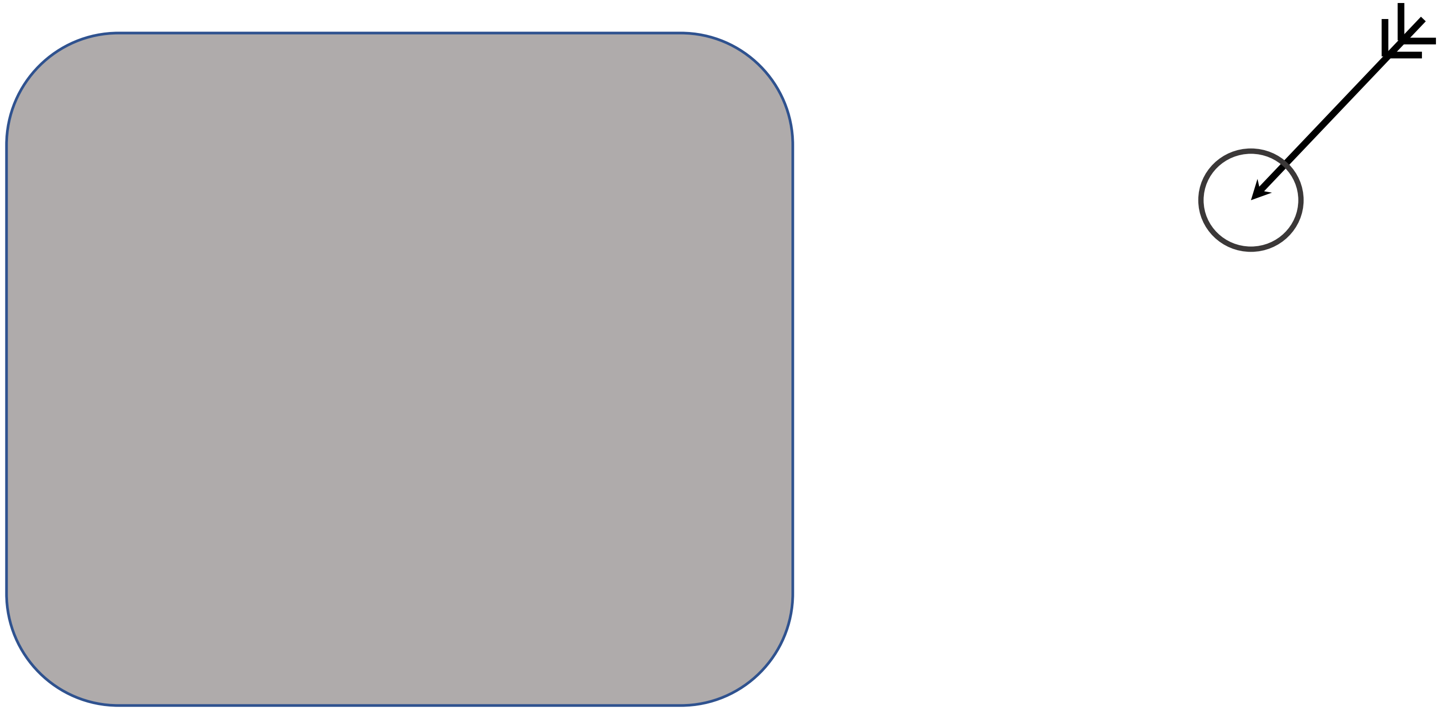
- We are tasked with identifying where on a target archery board the bullseye is placed. But all we can see is the back of the target and where the arrows puncture through as a marksman fires at it.

Back of target: What we see.

Front of target?

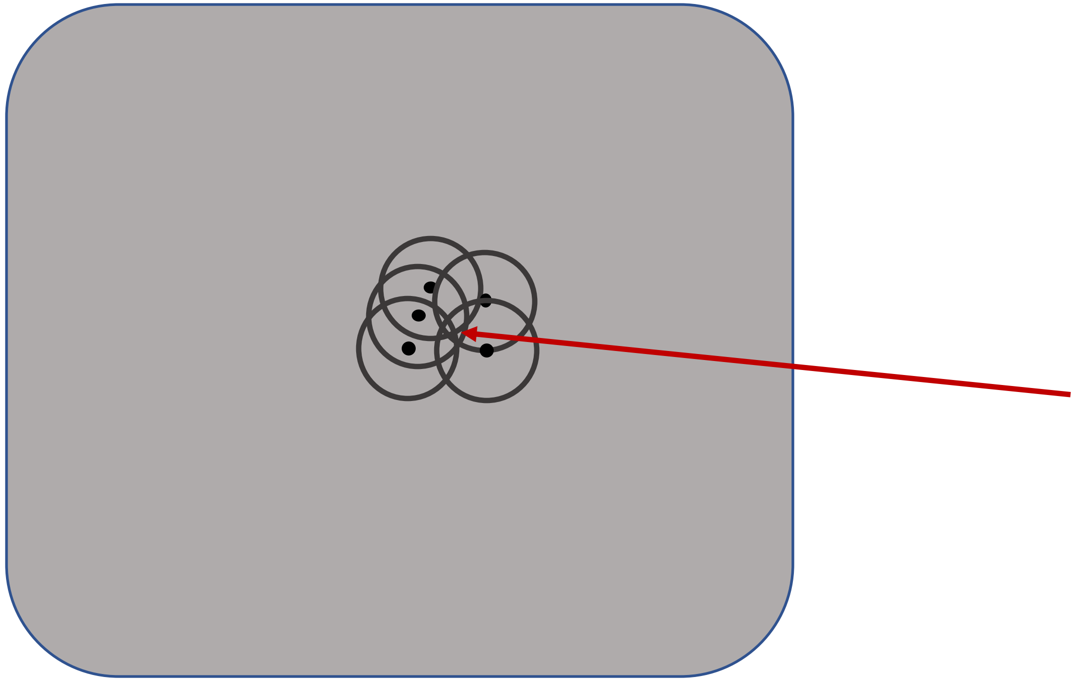
Archery Example Continued

- We are told that a marksman is firing at the target and when they fire, they are always within 10 centimeters of the target 95% of the time.



Archery Example Continued

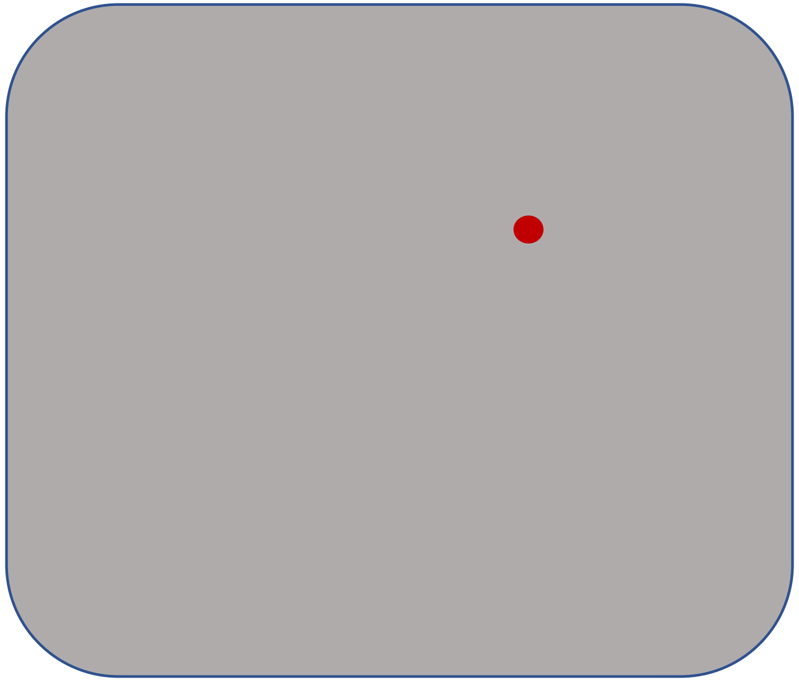
- Here are the archer’s first 5 shots with a 10-cm radius around it.
- A frequentist observed the most likely point as the point with the closest distance to the points as the best guess.



Most likely estimate.

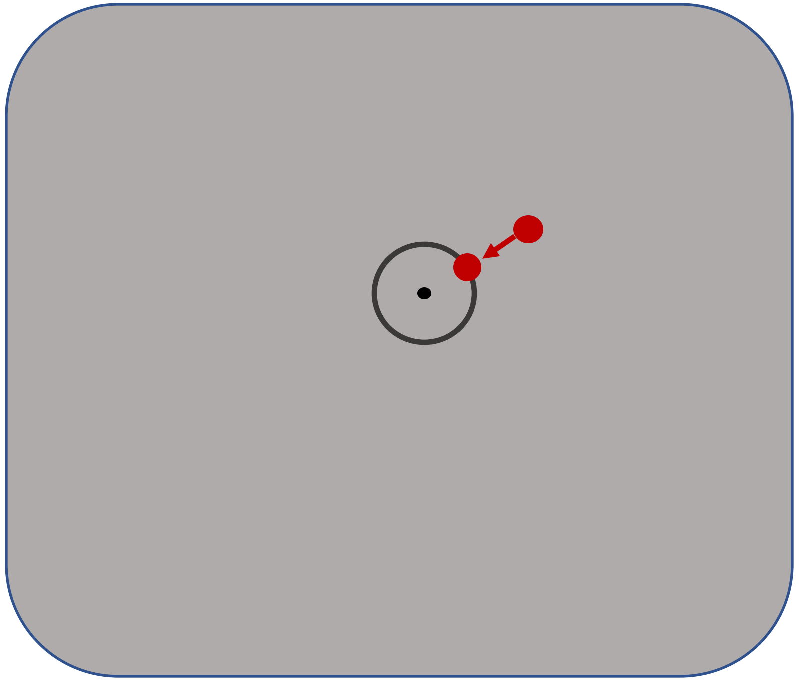
Archery Example

- A Bayesian Approach is to *start* with a “Prior”, or a previous belief of where the target is. (Red Point)



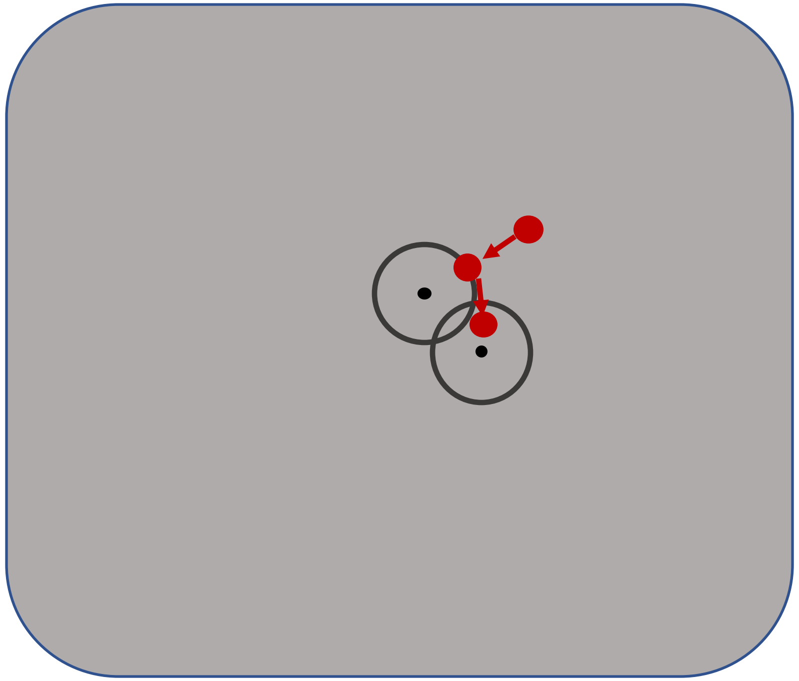
Archery Example

- As the next arrow fires, we update our prior via the posterior distribution.
- We continue in this fashion until our final target is chosen.



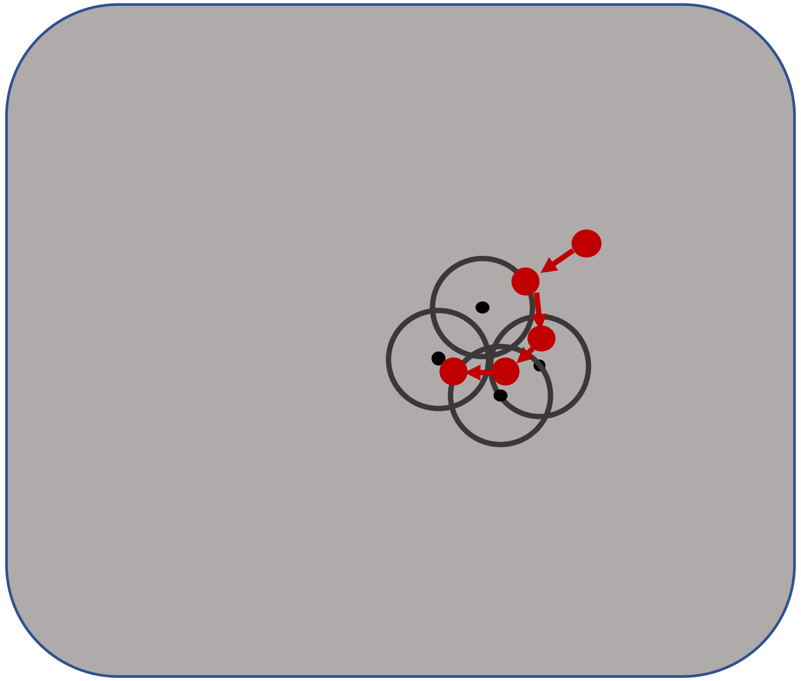
Archery Example

- As the next arrow fires, we update our prior via the posterior distribution.
- We continue in this fashion until our final target is chosen.



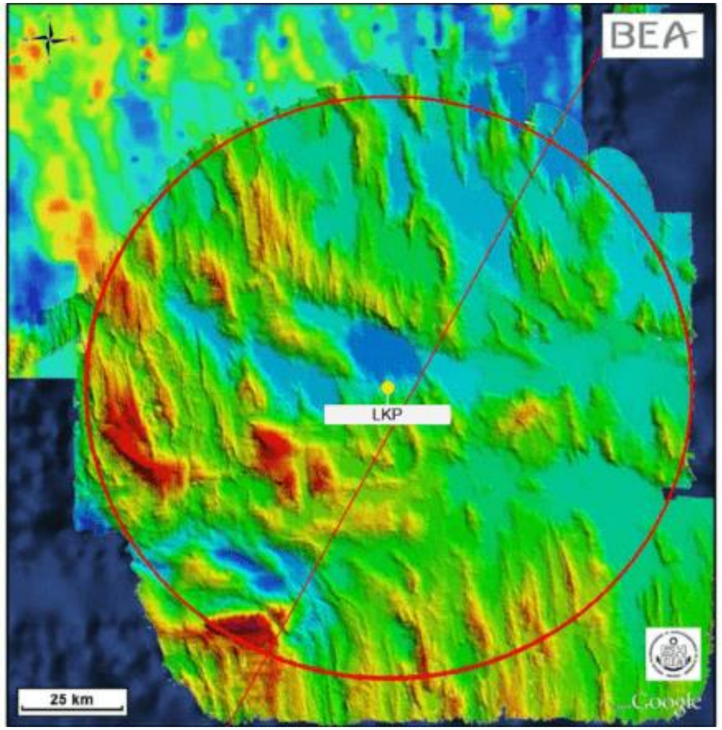
Archery Example

- As the next arrow fires, we update our prior via the posterior distribution.
- We continue in this fashion until our final target is chosen.



Real Life Example: Success in finding lost plane wreckage

- In practice, Bayesian inference has been used successfully to find lost planes. E.g. Air France 447.
- https://www.informs.org/ORMS-Today/Public-Articles/August-Volume-38-Number-4/In-Search-of-Air-France-Flight-447



discussion

True or False? The p-value is the probability that the null hypothesis is true.

A lot of students (and experienced scientists, minus statisticans) get the answer to the above question wrong. The reason they do is related to confusing Frequentist and Bayesian paradigms.

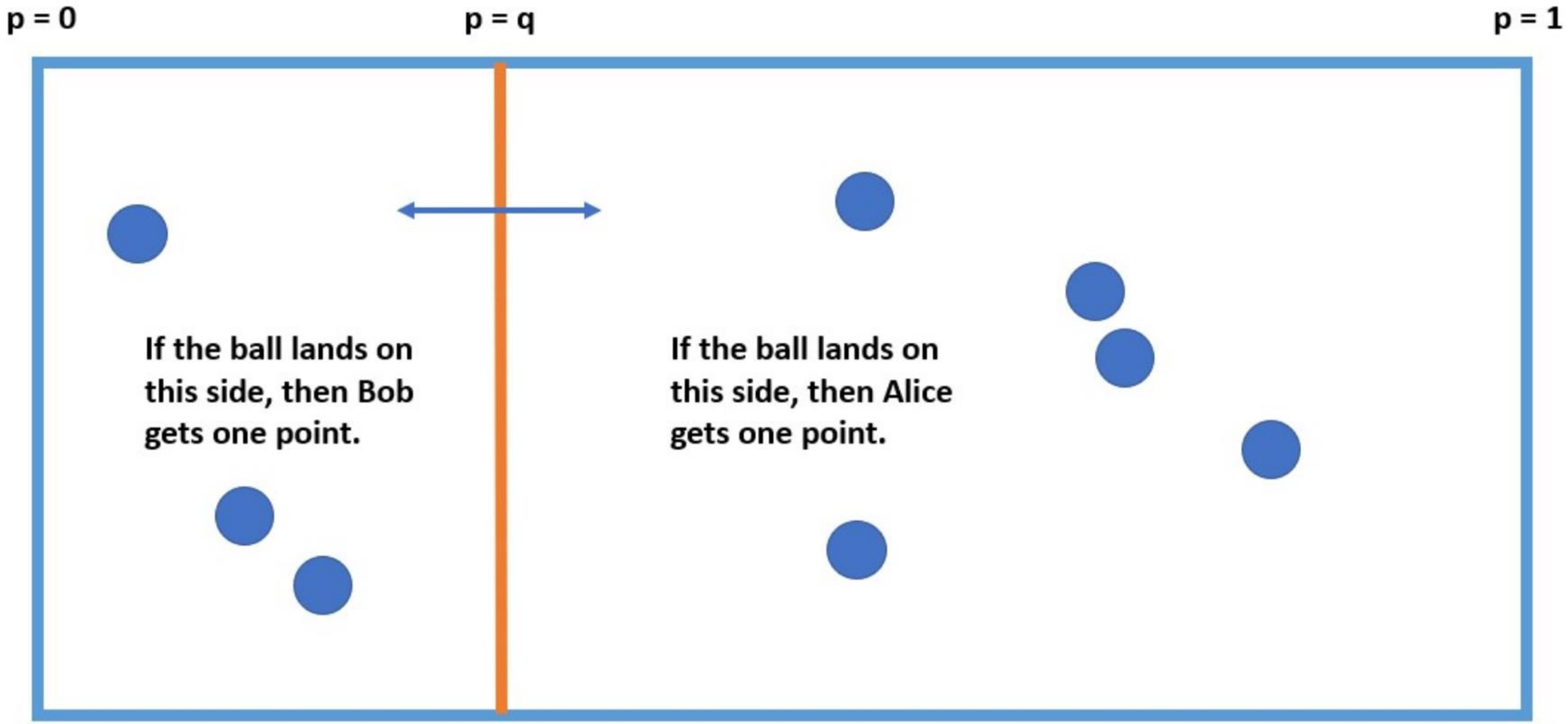
Take a moment to think about what the answer is, and why.

the billiards game

You roll a ball on a billiard table and mark its position on the table length-wise (across only one dimension). Alice and Bob take turns rolling balls and look at where they land in relation to where the first ball landed: if they land on the left side of the table Alice gets a point, otherwise Bob gets a point.

- the first to reach 6 points wins the game
- Alice and Bob don't know where the first ball landed, but they know where the other balls landed and what the score is
- denote the **current score** with D (as in *data*): as it stands we have Alice 5 and Bob 3 (shown in the next page)

Question: What is the probability of Bob winning the game? Note that for this to happen the ball needs to land on Bob's side the next three rounds.



p = 0

p = q

p = 1

If the ball lands on this side, then Bob gets one point.

If the ball lands on this side, then Alice gets one point.

the Frequentist approach

Let B be the event that Bob wins. Let D be the data, which is the current score. We want to know $P(B|D)$.

- Based on the data, the most likely value for p is $3/8$, i.e. the average of the data so far: $\hat{p} = 3/8$.
- We can write $P(B|D)$ as $P(B|D, p = 3/8)$. Since p is assumed to be fixed, there's no harm in conditioning on p.
- Since Bob needs to win the next three rounds, then he has a probability of $\hat{p}^3 = (3/8)^3 \approx 0.05$ of winning the game.

the Bayesian approach

- treat p as a random variable from a uniform distribution between 0 and 1
- rewrite $P(B|D)$ as

$$P(B|D) = \frac{P(B|D)P(D)}{P(D)} = \frac{\int_0^1 P(B|D,p)P(D|p)P(p)dp}{\int_0^1 P(D|p)P(p)dp}$$

  since we don't know the value of p, we just average out (integrate) over all possible values between 0 and 1

- plug in the probabilities and solve (analytically or numerically):

$$P(B|D) = \frac{\int_0^1 p^3(1-p)^5 p^3 dp}{\int_0^1 (1-p)^5 p^3 dp} = \frac{\int_0^1 (1-p)^5 p^6 dp}{\int_0^1 (1-p)^5 p^3 dp} \approx 0.09$$

Bayesian inference in practice

1. Identify data relevant to the research question.
2. Define a descriptive model for the data. For example, pick a linear model formula.
3. Specify a prior distribution for the parameters.
4. Let the Bayesian packages like `pymc` and `pystan` run through the computation to estimate the posterior distribution. Once we have the posterior distribution, we can generate parameters from it for simulations if we need to.
5. Update the posterior distribution if more data is observed. This is key! The posterior of a Bayesian model naturally updates as more data is added, a form of **online learning**.

A Coin Flip: Frequentist Perspective.

- We will flip a coin N times. We count the number of heads and want to estimate the p(H). E.g. if it is a fair coin, we would expect (with enough trails) that we would estimate p(H)=0.5.
- Frequentist probability:
  - Most likely answer would be:

$$p(H) = \frac{n(H)}{N}$$

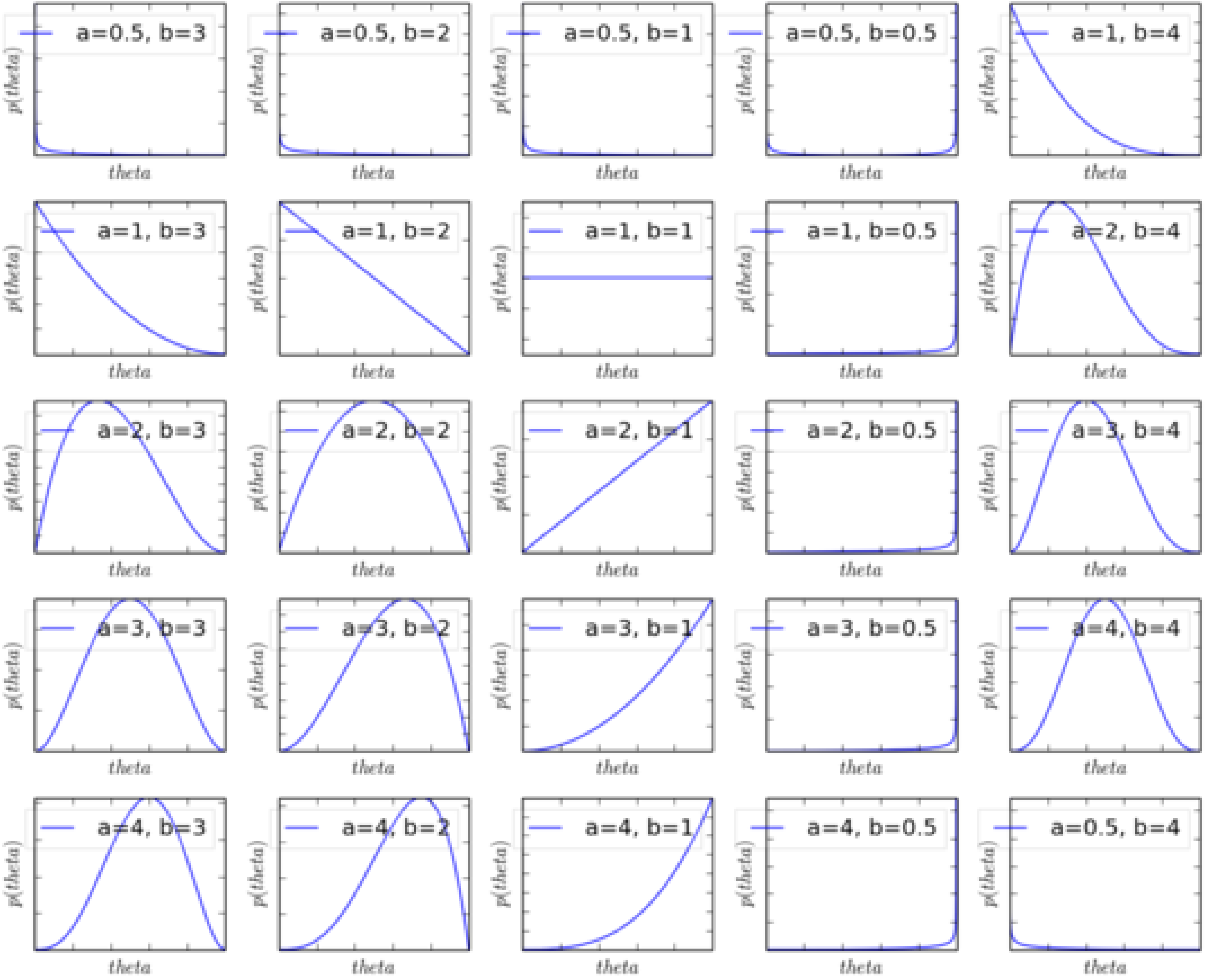
A Coin Flip: Bayesian Perspective.

- We will flip a coin N times. We count the number of heads and want to estimate the p(H). E.g. if it is a fair coin, we would expect (with enough trails) that we would estimate p(H)=0.5.
- Bayesian:
- We need to define a prior probability for the estimation of p(H).
- The best choice for N-coin flips is a Beta Distribution:

$$f(X) = x^{a-1}(1-x)^{b-1}$$

- a, b are constants that define the distribution (similar to how the mean and variance define different normal).
- X is only defined between 0 and 1.
  - This is what we want the range of p(H) to be.

Refresher on the Beta Distribution, B(x|a,b)



Bayesian Estimation of a Coin Flip Probability

- P(parameters|data)=P(data|parameters)*P(parameters)/P(data)
- Prior:

$$f(x) = x^{a-1}(1-x)^{b-1}$$

- After we choose a prior, we compute the posterior:

$$Posterior = Likelihood\,\frac{Prior}{P(data)}$$

- Have a problem estimating P(data). So, we write it out:

$$Posterior = Likelihood\,\frac{Prior}{P(data|all\ parameters)}$$

$$Posterior = Likelihood\,\frac{Prior}{\sum P(data|theta)}$$

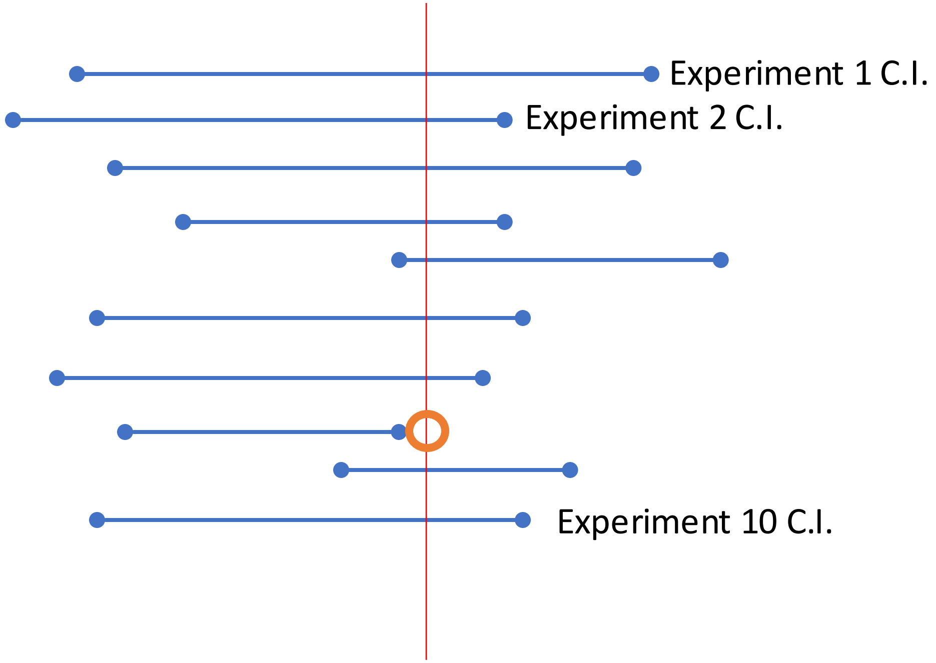
Bayesian Estimation Continued

- We only had one parameter to estimate for the coin flip example: p(H).
- We can create a grid to check for p(H), from 0.01, 0.02, 0.03, …, 0.99.
  - We iterate through all these thetas and calculate the p(data).
- What if we had several parameters?
  - If we had just 6 parameters with a length 100 grid to check for each: 100^6 = 1 TRILLION combinations to check.
- Maybe we don't have to sample *everything*. Just enough points to understand and estimate the distribution of how p(data) behaves under the 6 parameters.

- (Something to think about for next class…)

Confidence Intervals (Frequentist View)

> A frequentist 90% **confidence interval** means that with a large number of repeated samples, 90% of such calculated confidence intervals would include the one true value of the parameter.

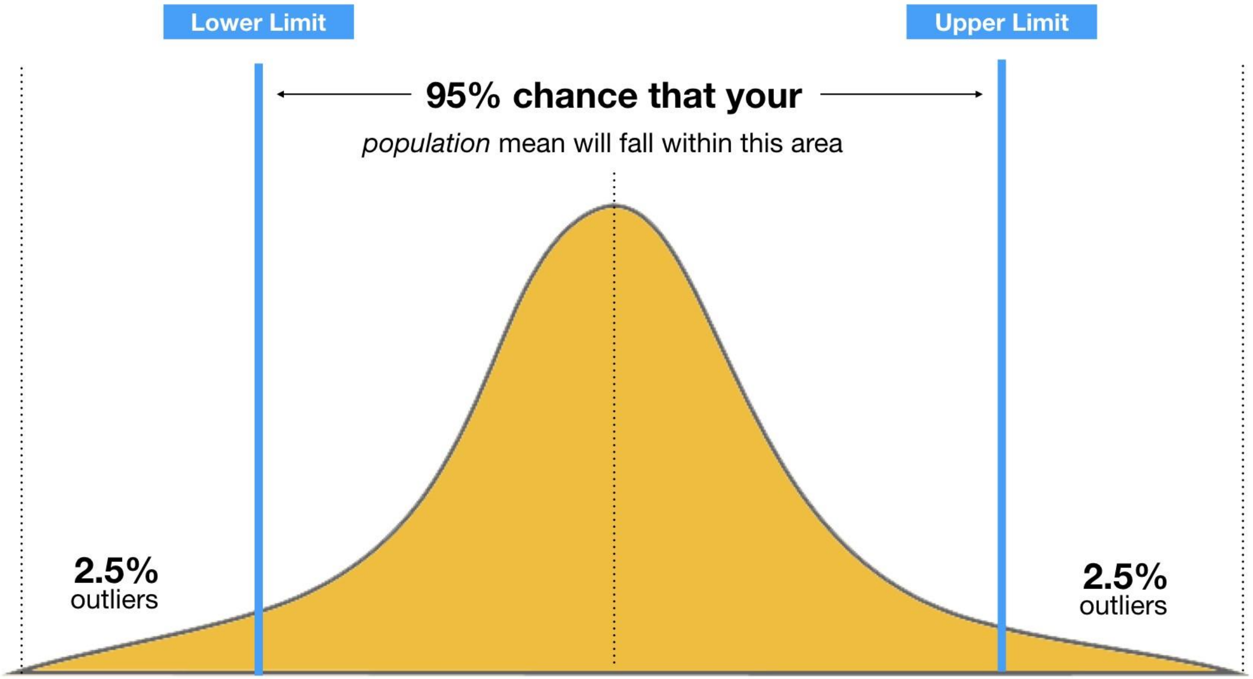


<u>True Population Value-</u>
This is some fixed value,
unknown to us.

Credible Intervals (Bayesian View)

- The true population value comes from a distribution. This interval covers 95% probability of that distribution, *given* the evidence provided by the observed data.
- This means that 95% of the time, our C.I. will cover the true value.

If we look at a sampling distribution, we can approximate a C.I. by taking the 2.5% and 97.5% percentiles for a 95% C.I.



discussion

A **confidence interval (CI)** is an closed interval around the **point estimate** (like a sample mean). So you can think of at as a **range estimate**, where the higher the **condfidence level** the wider the range. So a 99% CI is wider than a 95% CI, because we need to widen the range in order to speak with more confident.

Here are two interpretations for the confidence interval. Which interpretation relies the Frequentist paradigm and which relies on the Bayesian paradigms?

- 95% of the time, the true value will lie in this interval
- if we sample from the data and recompute the interval, 95% of such intervals will contain the true value

notebook time

we return to the lecture later

the end