Statistics for Life Sciences - Bayesian Statistics

Nathan Harmston

2022-04-05

Inverse problems

given P(A|B) what is P(B|A)?

reverse conditional probabilities

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

We don't do Bayesian calculations intuitively

Question

Only a tiny fraction (0.1%) of the people have a disease D. A test for this disease is highly accurate but not quite perfect. It correctly identifies 99% of patients with the disease but also incorrectly concludes that 1% of the noninfected samples have the disease. When this test identifies a blood sample as having HIV present, **if you have a positive result what is the chance that you have the disease**?

- P(D) - probability of having the disease -
- P(T) - probability of having a positive test
  - $P(b) = \Sigma P(a_i) x P(b|a_i)$ - law of total probability
- P(D | T) = what we want to find - probability of disease given a positive test
- P(T | D) = probability of a positive test given you have the disease

$$P(D|T) = 0.09$$

the interpretation of the result depends on the fraction of the population that has the disease

- statistical inference
- spam filters
- Bayes theory is to theory of probability what the Pythagorean theorem is to geometry - *some smart statistician*
- allows you to invert probabilities
- *Bayesian statistics*

Bayes rule

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

$P(A|B)$ = probability (A) given (B)
$P(B|A)$ = probability of (B) given (A)
$P(A)$ = probability of (A)
$P(B)$= probability of (B)

$P(A|B)$= posterior probability
$P(B|A)$ = likelihood
$P(A)$= prior probability
$P(B)$= marginal likelihood

Monty Hall Problem

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Monty Hall Problem

1

2

3

Monty Hall Problem

1

2

3

Monty Hall Problem

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat (never the car). He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

so would you switch or not?

Monty Hall Problem

3 doors...

- C = number of the door with the car
- S = number of the door selected
- O = number of the door opened

Monty Hall Problem

so ...

$$S = 1$$

1	2	3	Result if stick	Result is switch
G	G	C	G	C
G	C	G	G	C
C	G	G	C	G

so switching wins 2/3 of the time

The contestant

the contestant's choice and the door with the car are independent of each other, so

$$P(C = c) = 1/3 \; \forall c$$
$$P(C = c|S = s) = P(C = c) = 1/3 \; \forall c$$

The host

the choice by the host is not independent of the car or the contestants choice

$$P(O = o|C = c, S = s) = \begin{cases} 0, & \text{if o=s,} \\ 0, & \text{if o=c,} \\ 1/2, & \text{if o} \neq \text{s and s=c,} \\ 1, & \text{if o} \neq \text{c and o} \neq \text{s and s} \neq \text{c,} \end{cases}$$

using Bayes Formula

$$P(C = c|O = o, S = s) = \frac{P(O=o|C=c,S=s)P(O=o|C=c,S=s)}{P(O=o|C=c,S=s)}$$

and ...

$P(O = o|C = c, S = s)$ can be written as

$$\sum_{c=1}^{3} P(O = o|C = c, S = s)P(C = c|S = s)$$

Monty Hall Problem

$$P(O=o|C=c,S=s)=\begin{cases}0, & \text{if o=s,}\\ 0, & \text{if o=c,}\\ 1/2, & \text{if o} \neq \text{s and s=c,}\\ 1, & \text{if o} \neq \text{c and o} \neq \text{s and s} \neq \text{c,}\end{cases}$$

what is $P(C=3|O=2,S=1)$?

the host is providing us with additional information by opening a door and this is altering the resulting probabilities

`https://web.archive.org/web/20130121183432/http://marilynvossavant.com/game-show-problem/`

What is Bayesian statistics?

Sean R Eddy

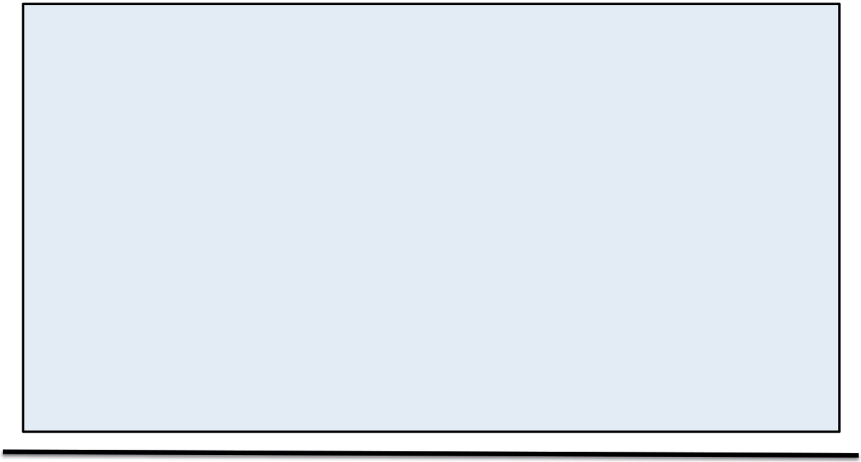
There seem to be a lot of computational biology papers with 'Bayesian' in their titles these days. What's distinctive about 'Bayesian' methods?

Sean Eddy 2004

The table game

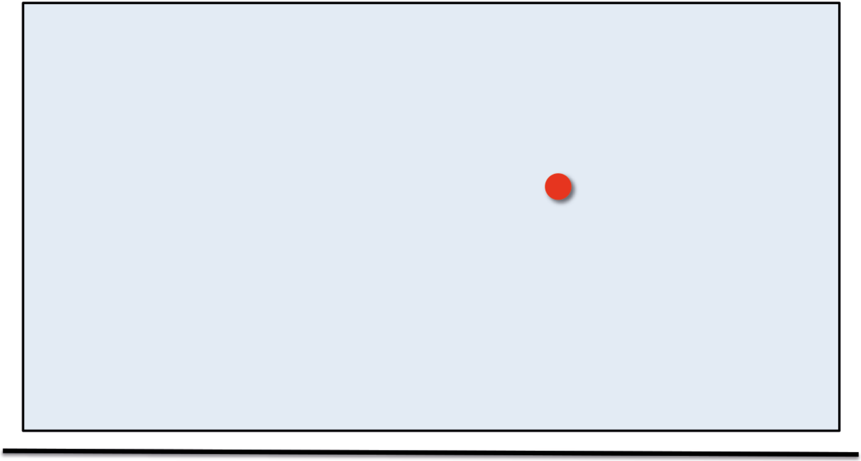
- Nathan and Ajay give up on science, turn to gambling, and go to a casino
- they join the table game where they are seated and so they can't see the table
- the house rolls a ball onto the table and everything to the left of the ball is Ajay's and everything to the right is Nathans's

the table game



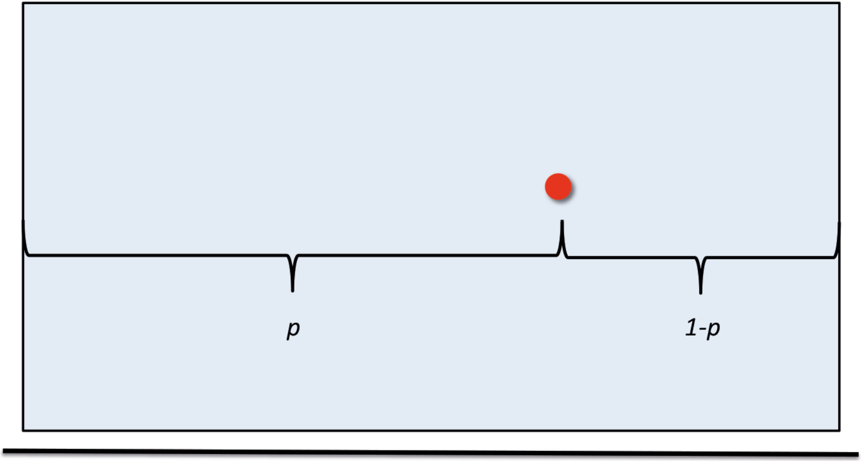
Ajay - 0; Nathan - 0

the table game



Ajay - 0; Nathan - 0

the table game

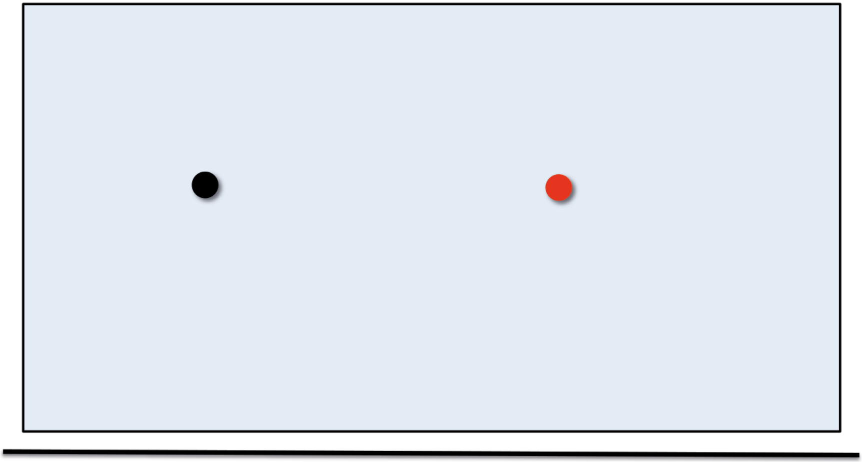


Ajay - 0; Nathan - 0

- The house rolls additional balls onto the table.
- If the balls land on the left, Nathan gets a point and if it lands to the right, Ajay gets a point.
- The first person to reach 6 points wins.

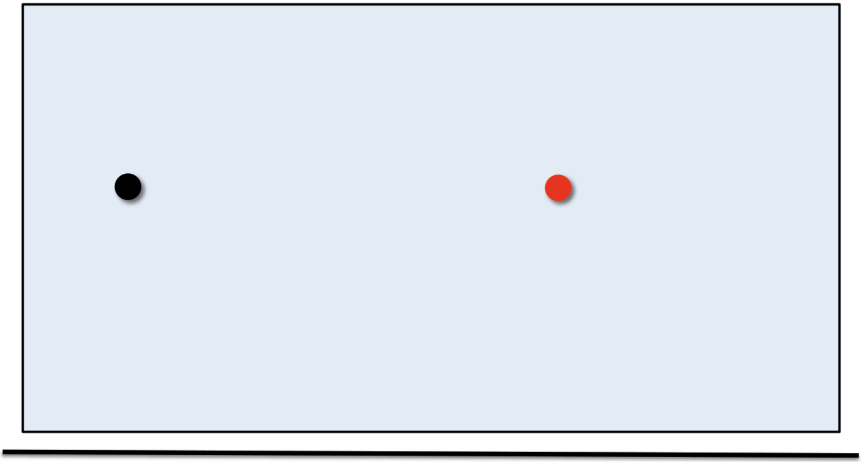
Ajay - 0; Nathan - 0

the table game



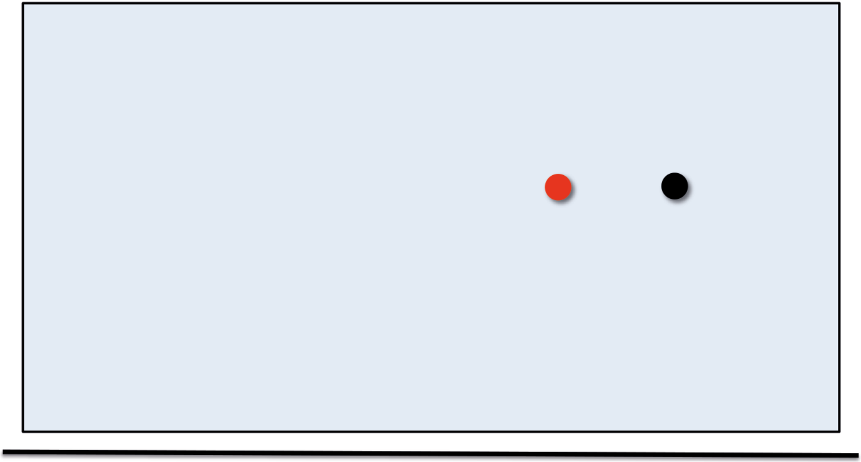
Ajay - 1; Nathan - 0

the table game



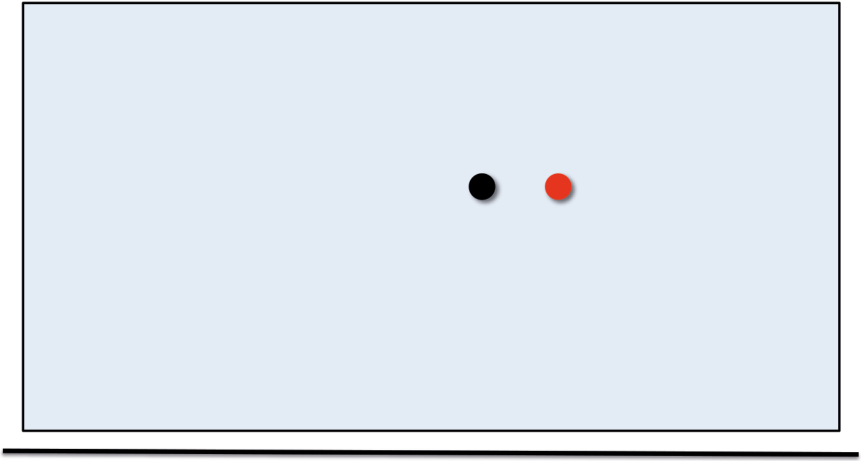
Ajay - 2; Nathan - 0

the table game



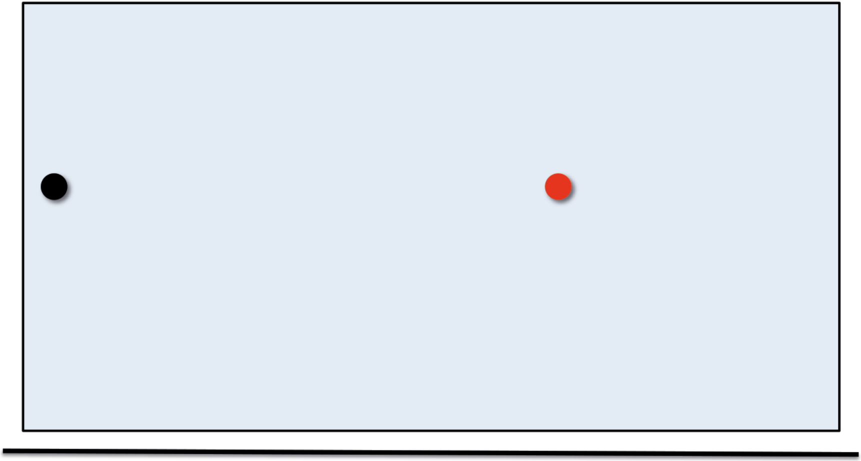
Ajay - 2; Nathan - 1

the table game



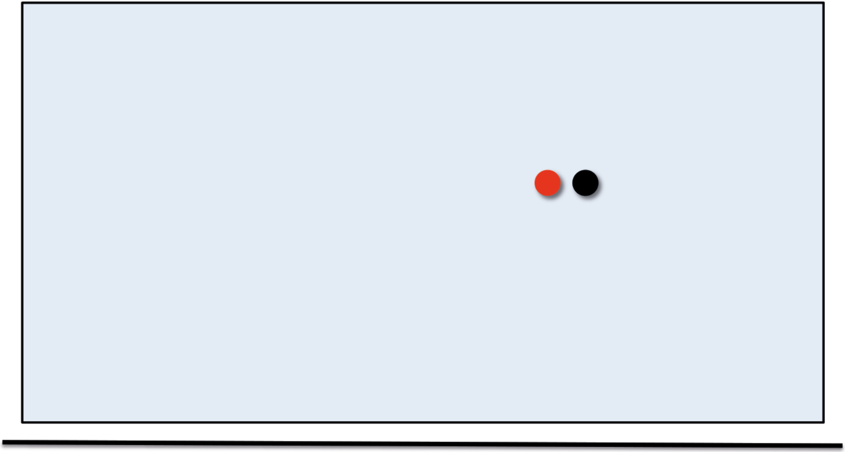
Ajay - 3; Nathan - 1

the table game



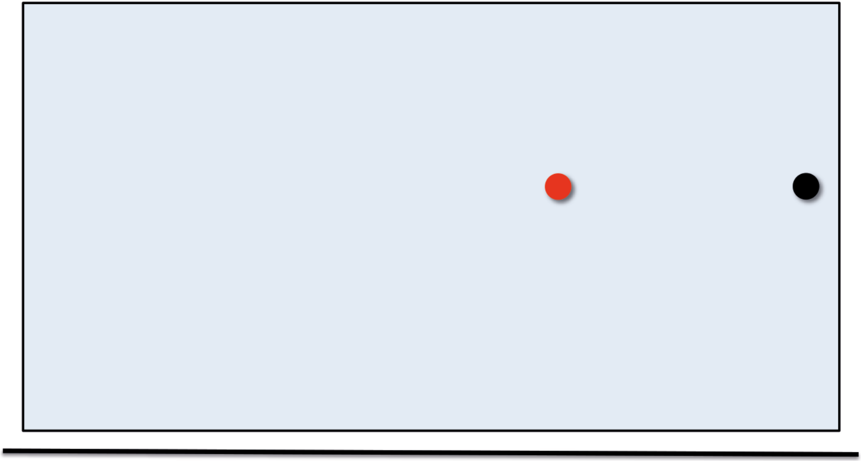
Ajay - 4; Nathan - 1

the table game



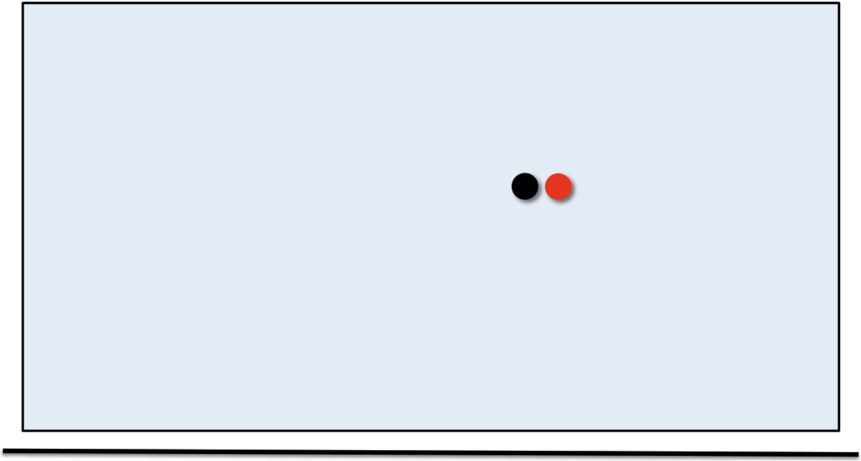
Ajay - 4; Nathan - 2

the table game



Ajay - 4; Nathan - 3

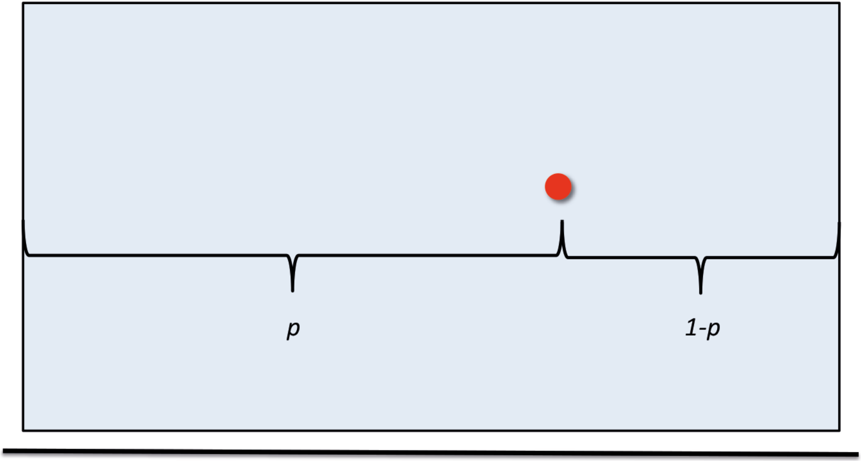
the table game



Ajay - 5; Nathan - 3

Now for a bet?

- Ajay now has 5 and Nathan has 3?
- What is the probability that Nathan will win?
- What is the probability that Ajay will win?
- What are the odds that Ajay will win?



Ajay - 5; Nathan - 3

- probability that Nathan will win is $(1-p)^3$
- probability that Ajay will win is $1-(1-p)^3$
- what are the odds that Ajay will win?

- probability that Nathan will win is $(1-p)^3$
- probability that Ajay will win is $1-(1-p)^3$
- what are the odds that Ajay will win?

$$\text{odds} = \frac{1-(1-p)^3}{(1-p)^3}$$

- probability that Nathan will win is $(1-p)^3$
- probability that Ajay will win is $1-(1-p)^3$
- what are the odds that Ajay will win?

$$\text{odds} = \frac{1-(1-p)^3}{(1-p)^3}$$

but wait....

we don't know p, we can only estimate it

the Frequentist approach

$$p = \frac{5}{8}$$

$$1 - p = \frac{3}{8}$$

the odds of Ajay winning are ....

the frequentist way of thinking of about probabilities is as the probability of an event happening

large of large numbers etc.

$$p = \frac{5}{8}$$

$$1 - p = \frac{3}{8}$$

the odds of Ajay winning are ....

18:1

the frequentist way of thinking of about probabilities is as the probability of an event happening

large of large numbers etc.

The Bayesian approach

the ball could have ended up anywhere along the table with an equal probability (i.e. its uniform)

so the expectation of Nathan winning is the weighted average of $(1-p)^3$ over all possible values of p

$$E(\text{Nathan}) = \int_0^1 (1-p)^3 P(p|A=5, B=3)dp$$

the Bayesian way of thinking about probabilities is to represent a degree of belief

So lets invert things...

$$P(p|A=5,B=3) = \frac{P(A{=}5,B{=}3|p)P(p)}{\int_0^1 P(A{=}5,B{=}3|p)P(p)dp}$$

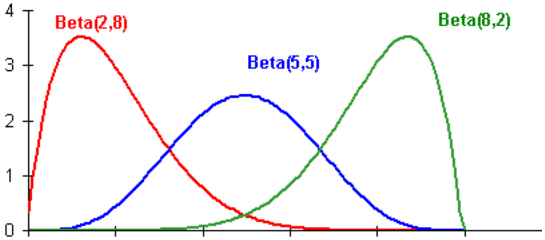
the likelihood is binomial and the prior is uniform (constant)

$$E(\text{Nathan}) = \frac{\int_0^1 p^5(1-p)^6 dp}{\int_0^1 p^5(1-p)^3 dp}$$

this has an analytical form - a beta integral which gives 1/11 for Nathans chance of winning and 10/11 for Ajay's chance of winning

the odds are 10:1 that Ajay will win

Beta distribution



The beta distribution is a often used to model percentages, proportions and probabilities

Can we simulate this?

```
set.seed(123)
winner = c()
for(i in 1:50000){
  p = runif(1, 0,1)
  tmp = sample(c("A", "N"), 8, replace=TRUE, prob=c(p, 1-p))
  if(sum(tmp=="A")==5){
    a = 5; n = 3
    while(TRUE){
      tmp = sample(c("A", "N"), 1, prob=c(p, 1-p))
      if(tmp =="A"){ a = a + 1 }
      if(tmp =="N"){ n = n + 1 }
      if(n == 6 || a == 6){ break }
    }
    if( a == 6 ){winner = c(winner, "A")}
    if( n == 6 ){winner = c(winner, "N")}
  }
}
```

I'm at a casino .... again

- so I'm tossing a coin yet again ?
- is it a fair coin or not?

so obviously I do some flips and count them .....

modified from
https://tinyheero.github.io/2017/03/08/how-to-bayesian-infer-101.html

Bayesian inference - is my coin fair?

Importance of the prior

The posterior is a balance between the prior and likelihood

- when there isn't a lot of data, the distribution is skewed towards the prior distribution
- when there is a lot of data, the distribution is skewed towards the likelihood distribution

so this is really useful it situations when you have limited amounts of data

I

have observed X positive tests and Y negative tests for COVID,

what is the proportion of individuals in my population who have COVID??

compute the posterior probability given by the number of successes and failures?

but this posterior is affected by your prior beliefs -

Bayes rule

$$P(\text{model}|\text{data}) = \tfrac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})}$$

$P(A|B)$ = probability of model (A) given the the data (B)
$P(B|A)$= probability the data (B) given the model (A)
$P(A)$= probability of the data
$P(B)$= probability of the model

why is this different to frequentist hypothesis testing?

in normal hypothesis testing - we've already assumed the null hypothesis is true - theres no way to include information on what the probability of it is

Benefits of Bayesian Statistics

- Explicitly write the inference problem
- Utilise prior information
- Can update the model as information is learned
- Infer uncertainties and missing data with probabilities

Difficulties with Bayesian statistics

Prior Specification

- Informative vs uninformative
- Subjective vs. biased

Numerical Integration

- Computationally intensive
- Requires methods like Gibbs sampling and Markov Chain Monte Carlo

Bayesian statistics

- Model based method with the incorporation of prior information
- Useful for biological systems with uncertain, noisy and/or missing data
- Can be computational expensive
- Widely used in the genetics and genomics fields, less so in the other ‘omics’

On Friday

- Model fitting / selection
- RSME, R-squared, AIC, BIC, cross-validation