Bayes

Presented by David John Baker
November 2019

FLATIRON SCHOOL

Bayesian Statistics

- Describe the relationship between Bayesian and Frequentist schools of thought
- Explain core concepts of Bayesian data analysis
  - Conjectures
  - Posterior distribution
  - Likelihood
  - Prior Distribution
  - Updating
- Review Classic Bayes Interview Q

Rev. Thomas Bayes

- 1701-1761
- Might not be Tom pictured here
- Died before his friend published his work

Rev. Thomas Bayes

- 1701-1761
- Might not be Tom pictured here
- Died before his friend published his work

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

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

Likelihood
How probable is the evidence given that our hypothesis is true?

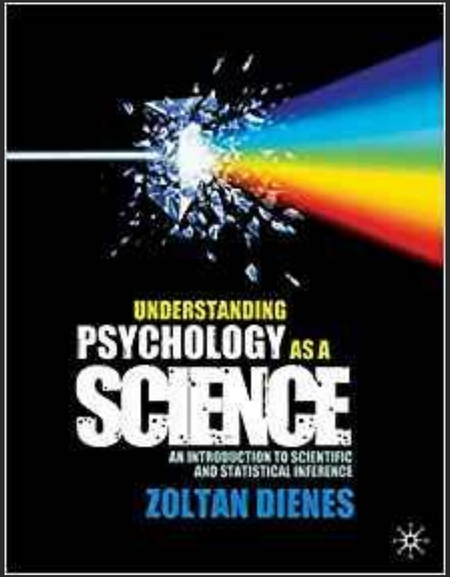
Prior
How probable was our hypothesis *before* observing the evidence?

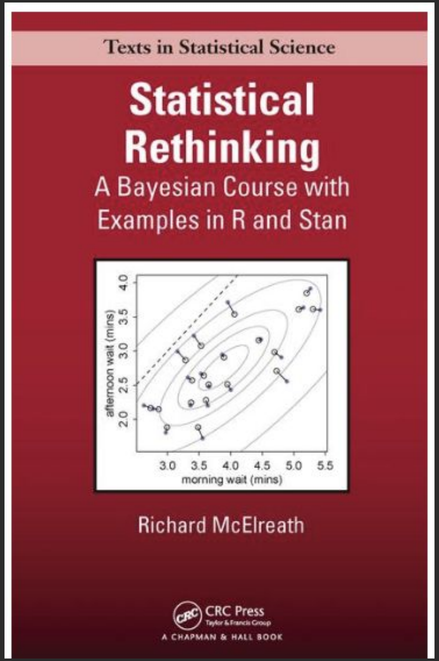
$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

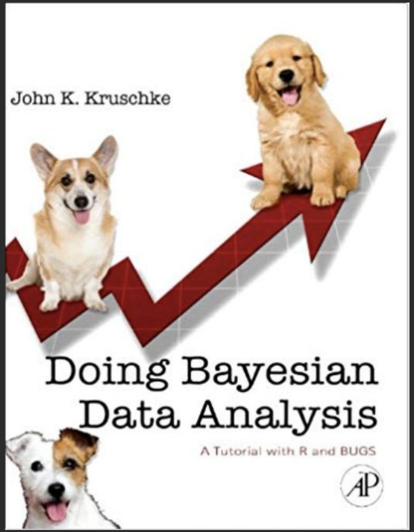
Posterior
How probable is our hypothesis given the observed evidence? (Not directly computable)

Marginal
How probable is the new evidence under all possible hypotheses?
$P(e) = \sum P(e \mid H_i)\, P(H_i)$

Bayesian Statistics







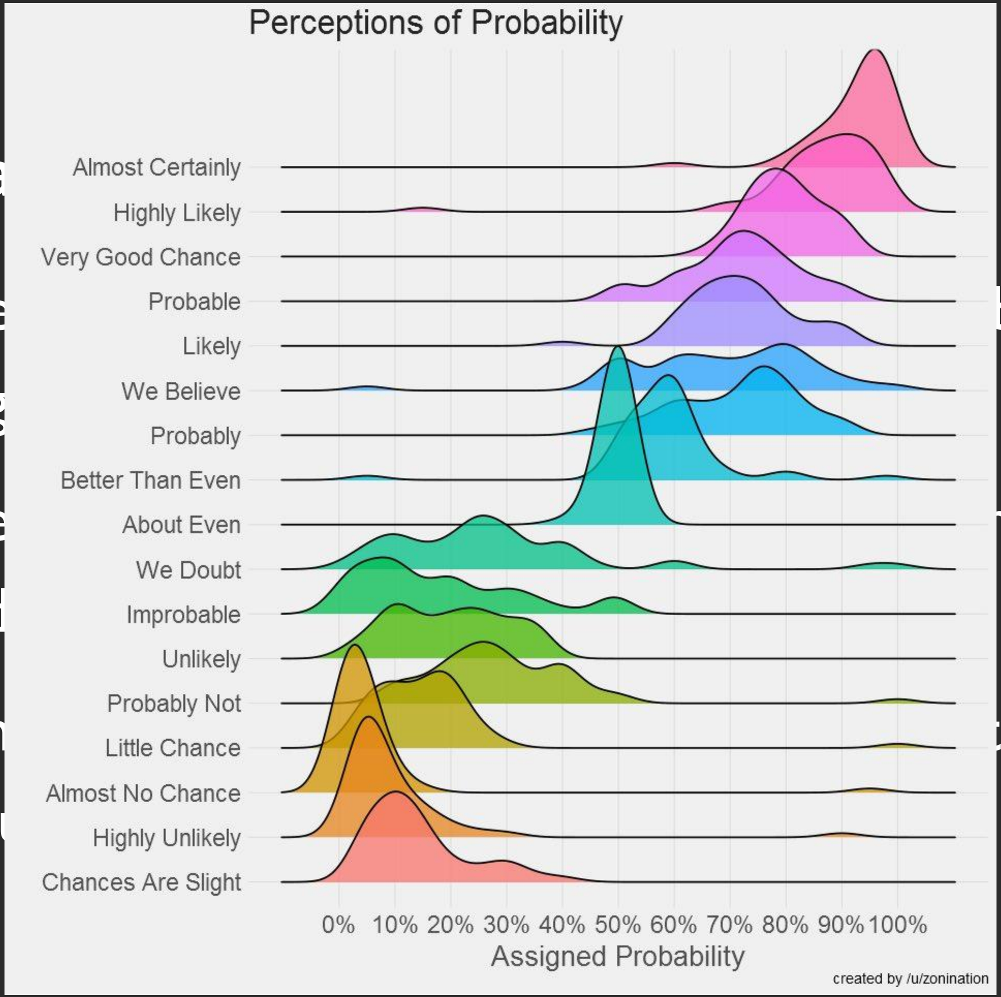
Bayesian Statistics

- Most people want statistics to tell them the probability of their hypothesis given their data
- Bayes will tell you the relative probability of your hypothesis given your data
- Frequentism will tell you the "objective" probability of your data give your (null) hypothesis

//

Bayesian Sta

- Most people … bility of their hypothesis g
- Bayes will te … hypothesis given your d
- Frequentism … y of your data give you



//

Bayesian Statistics

- Subjective probability is degree of conviction we have in hypothesis
- Subjective probabilities are in our mind, not in the world
- Data is the world, our models approximate the world

Bayesian Statistics

- Subjective probability is degree of conviction we have in hypothesis
- Subjective probabilities are in our mind, not in the world
- Data is the world, our models approximate the world
- DISCUSSION QUESTION: Using what you know thus far, write down two reasons to use and not to use both Bayesian methods vs frequentist methods

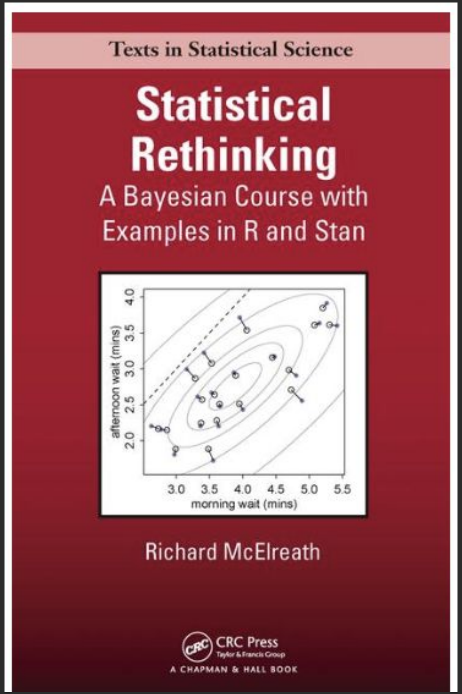
Concept Alert!

- Prior
- Likelihood
- Posterior Distribution
- Conjectures

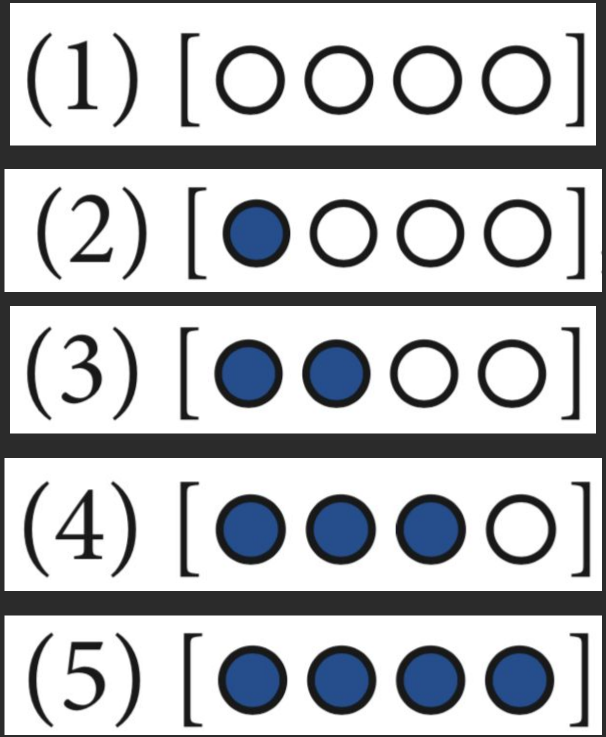
Building Intuitions

- **Marbles in a bag**

McElreath, R. (2018). *Statistical rethinking:*
A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.



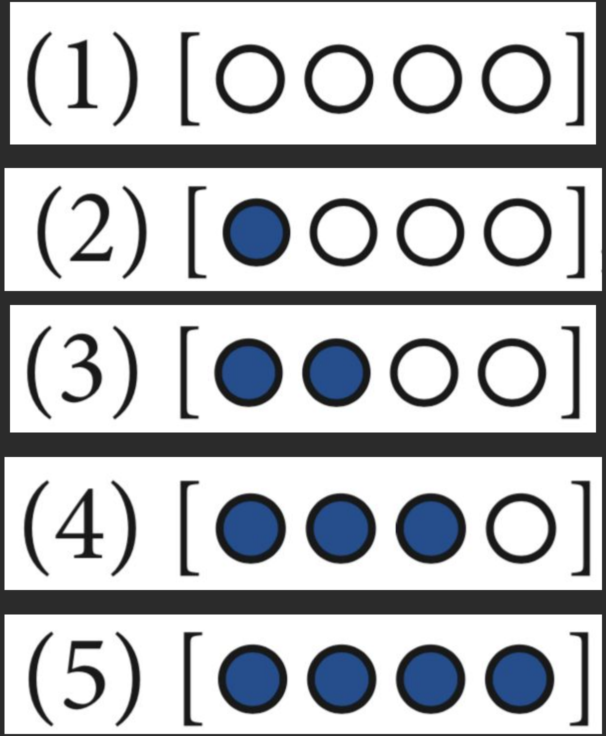
4 Marbles in a Bag, Two Colors — Five *Conjectures*



Scenario: What is most plausible?

Get to pull out three marbles
One at a time
Replace them

5 Marbles in a Bag, Two Colors — Five *Conjectures*

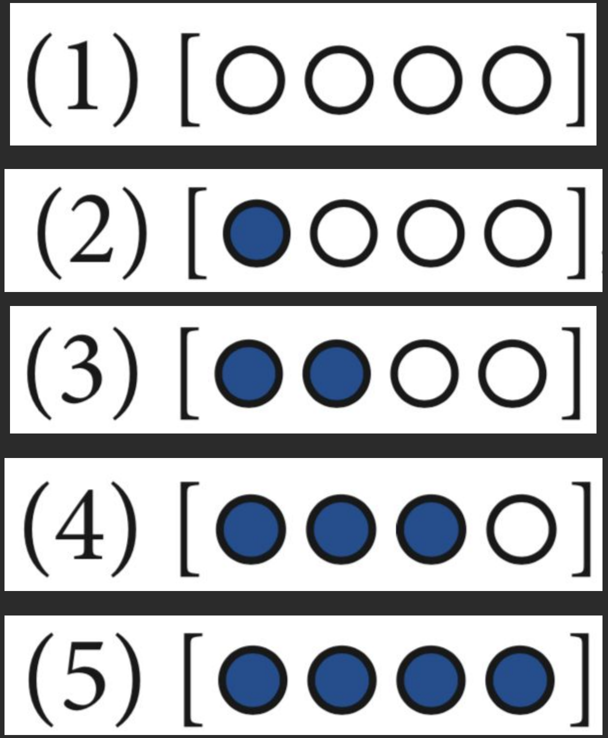


Scenario: What is most plausible?

Get to pull out three marbles
One at a time
Replace them

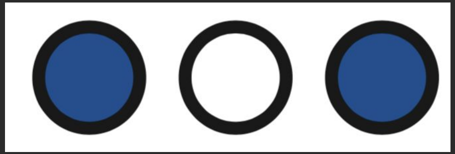
What do sample space and event
Space have to do with this scenario?

Marbles in a Bag — Five Conjectures

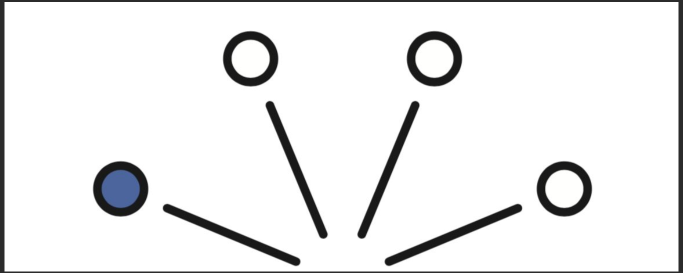


Scenario: What is most plausible?

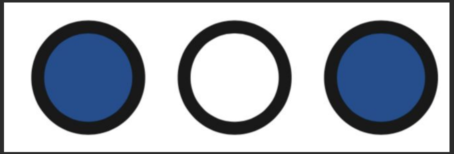
Get to pull out three marbles
One at a time
Replace them



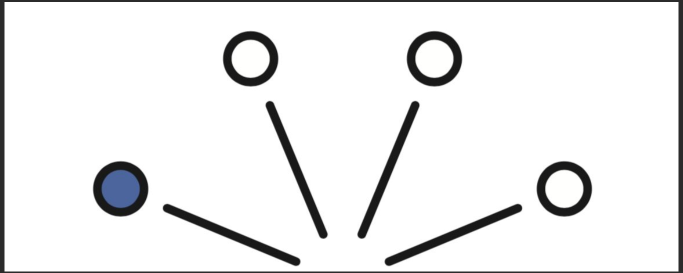
Single Conjecture?



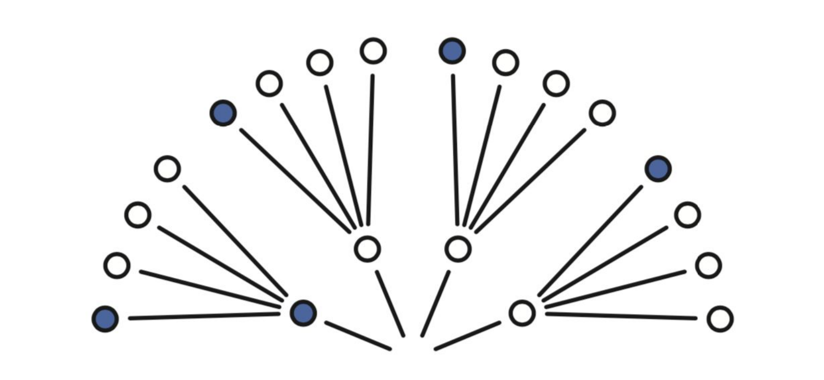
- First Draw?



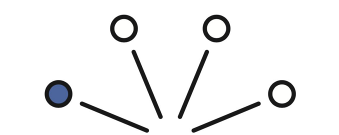
Single Conjecture?

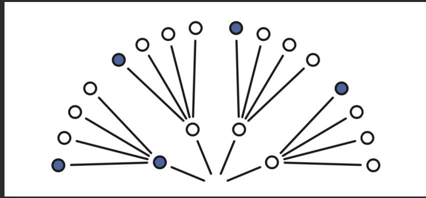


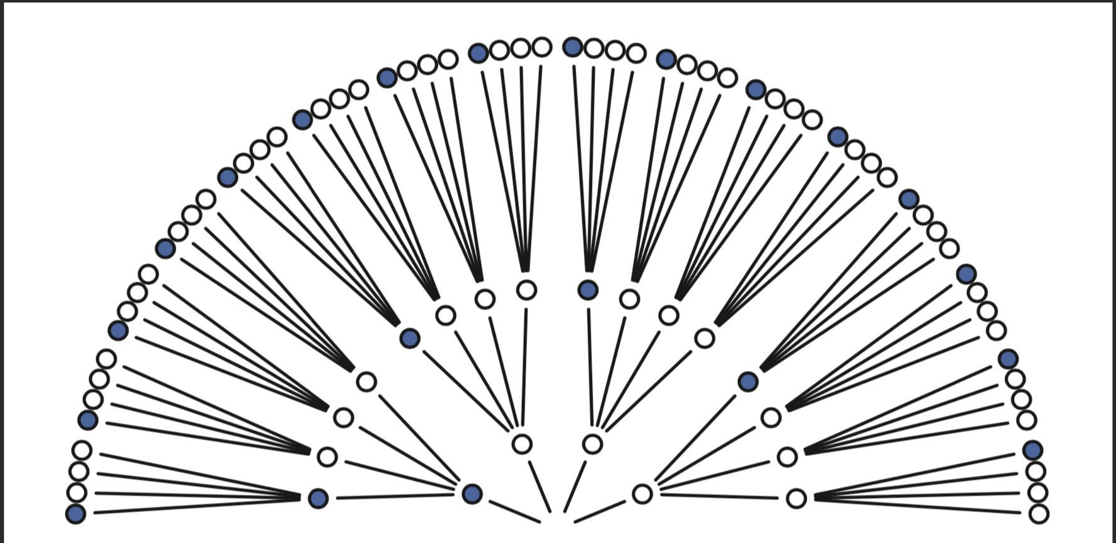
- Second Draw?



Single Conjecture?

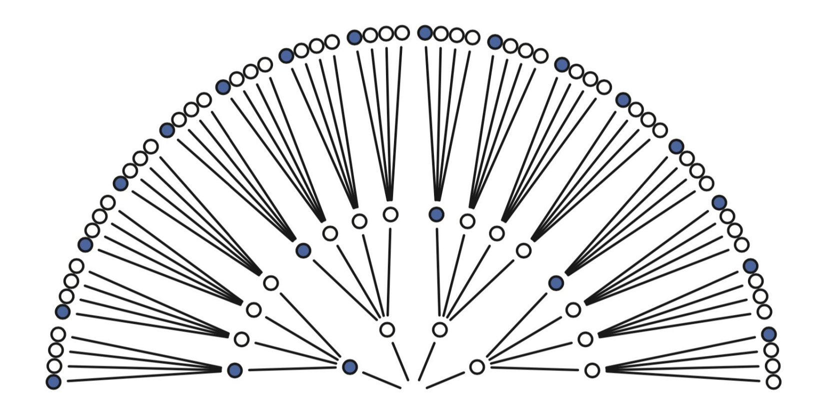


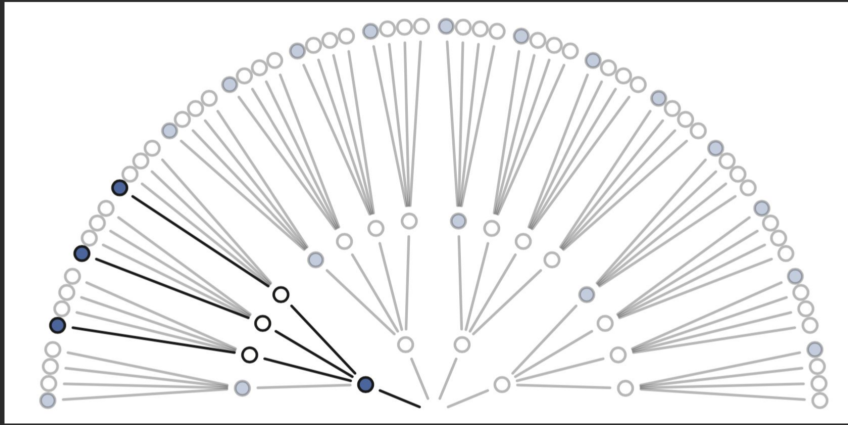




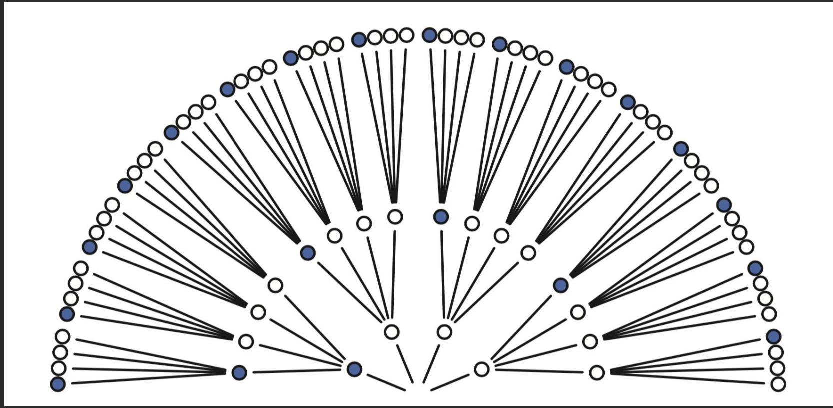
- Third Draw?

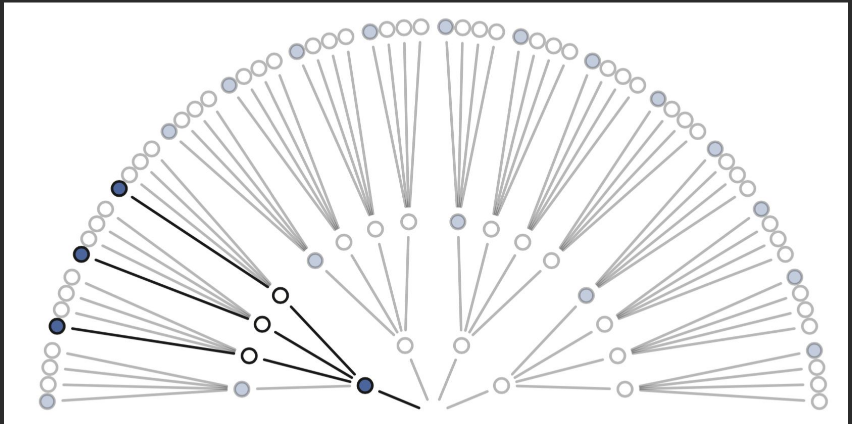
Single Conjecture?



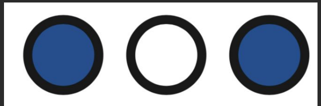


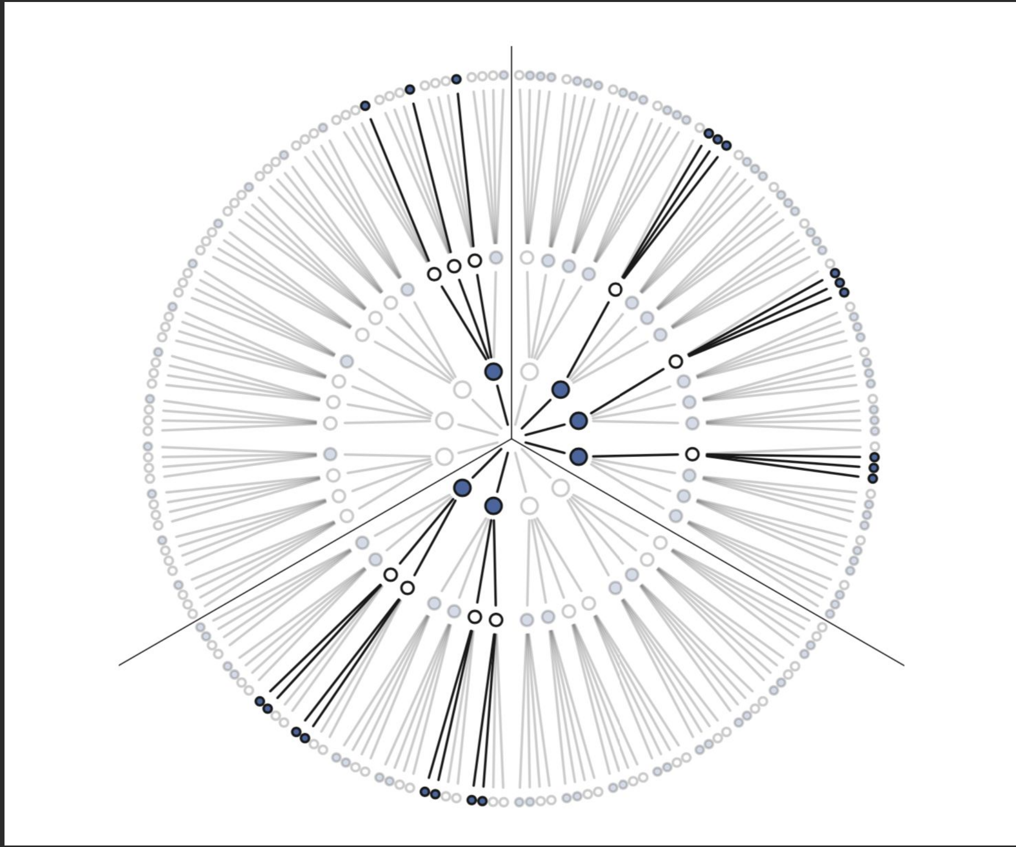
How many ways can produce ●○● ?





Conjecture	Ways to produce ●○●
[○○○○]	$0 \times 4 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$
[●●●○]	$3 \times 1 \times 3 = 9$
[●●●●]	$4 \times 0 \times 4 = 0$

How many ways can produce  ?



Draw a fourth blue?!?

Conjecture	Ways to produce ●	Previous counts	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	1	3	$3 \times 1 = 3$
[●●○○]	2	8	$8 \times 2 = 16$
[●●●○]	3	9	$9 \times 3 = 27$
[●●●●]	4	0	$0 \times 4 = 0$

Using Prior Information ??

Conjecture	Ways to produce ●	Previous counts	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	1	3	$3 \times 1 = 3$
[●●○○]	2	8	$8 \times 2 = 16$
[●●●○]	3	9	$9 \times 3 = 27$
[●●●●]	4	0	$0 \times 4 = 0$

Conjecture	Prior count	Factory count	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	3	3	$3 \times 3 = 9$
[●●○○]	16	2	$16 \times 2 = 32$
[●●●○]	27	1	$27 \times 1 = 27$
[●●●●]	0	0	$0 \times 0 = 0$

Counts to Probability

Conjecture	Ways to produce ●	Previous counts	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	1	3	$3 \times 1 = 3$
[●●○○]	2	8	$8 \times 2 = 16$
[●●●○]	3	9	$9 \times 3 = 27$
[●●●●]	4	0	$0 \times 4 = 0$

Possible composition	p	Ways to produce data	Plausibility
[○○○○]	0	0	0
[●○○○]	0.25	3	0.15
[●●○○]	0.5	8	0.40
[●●●○]	0.75	9	0.45
[●●●●]	1	0	0

Basic Bayes Terms

- A conjectured proportion of blue marbles, p, is usually called a **PARAMETER** value. It's just a way of indexing possible explanations of the data.
- The relative number of ways that a value p can produce the data is usually called a **LIKELIHOOD**. It is derived by enumerating all the possible data sequences that could have happened and then eliminating those sequences inconsistent with the data.
- The prior plausibility of any specific p is usually called the **PRIOR PROBABILITY**.
- The new, updated plausibility of any specific p is usually called the **POSTERIOR PROBABILITY**.

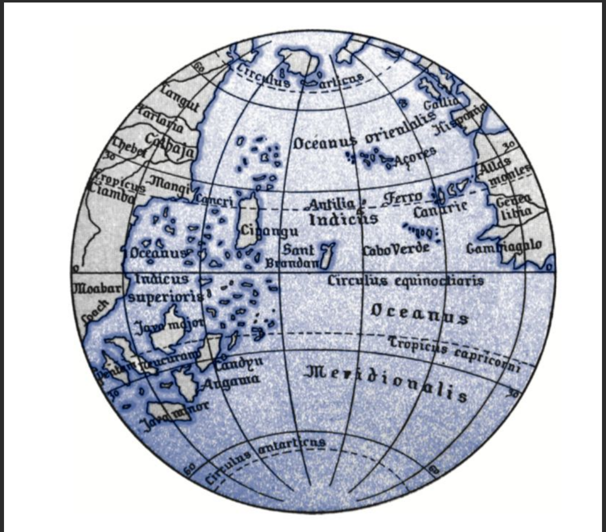
Please use the next 5 minutes to discuss how these four terms relate to the marbles in a bag example.

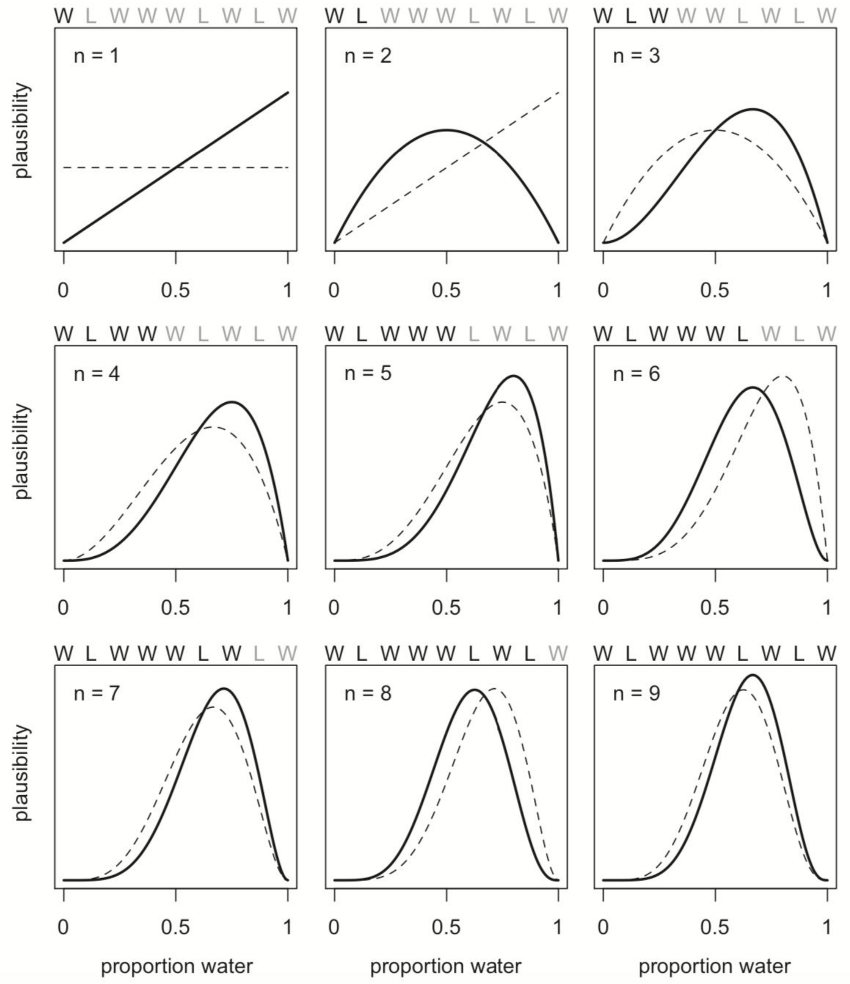
- A conjectured proportion of blue marbles, p, is usually called a **PARAMETER** value. It's just a way of indexing possible explanations of the data.
- The relative number of ways that a value p can produce the data is usually called a **LIKELIHOOD**. It is derived by enumerating all the possible data sequences that could have happened and then eliminating those sequences inconsistent with the data.
- The prior plausibility of any specific p is usually called the **PRIOR PROBABILITY**.
- The new, updated plausibility of any specific p is usually called the **POSTERIOR PROBABILITY**.

Updating

W L W W W L W L W

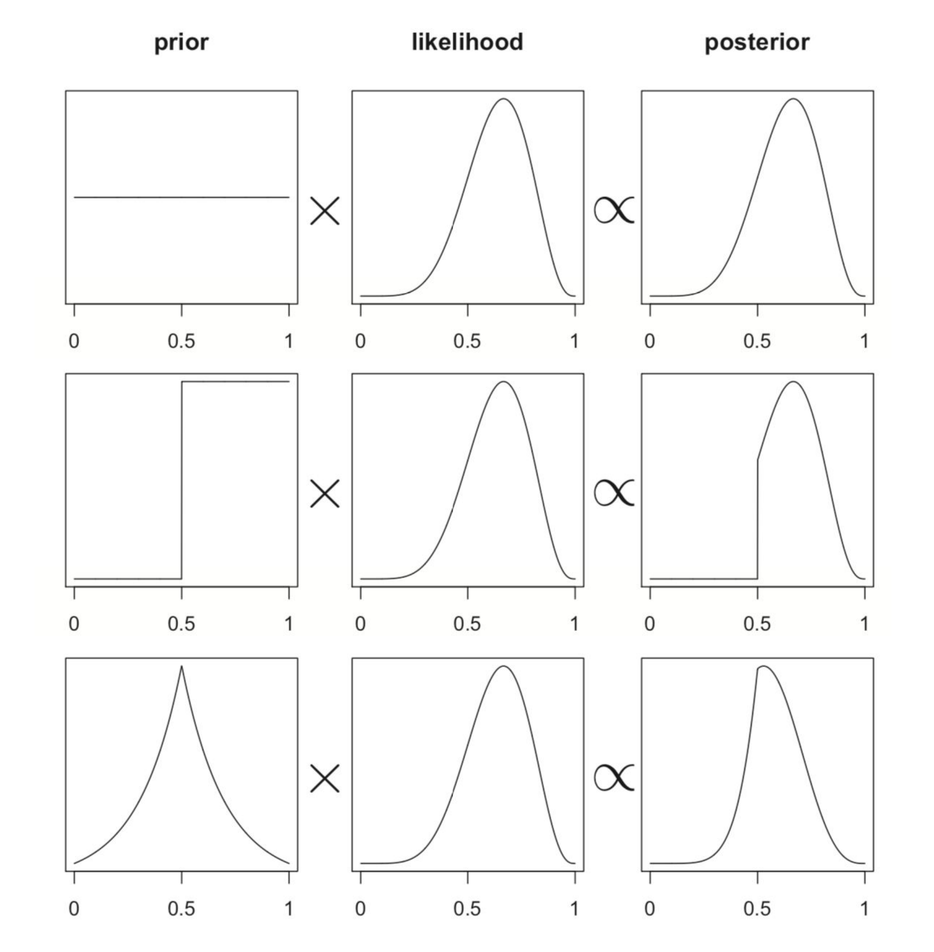
Imagine you have a globe and throw it
Want to determine % water
W means hand on water
L means hand on land




W L W W W L W L W
n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
n = 7
n = 8
n = 9
plausibility
0
0.5
1
proportion water

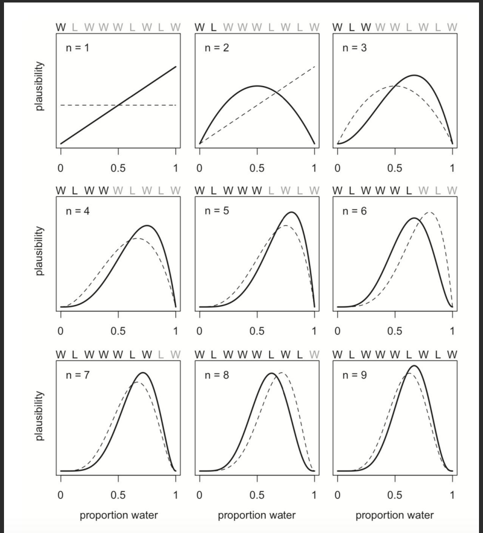
Bayesian Updating

- Given every new sample, new observations will change our posterior distribution
- Setting prior differently will make it harder/easier to move towards new posterior values


prior
likelihood
posterior
×
∝
0
0.5
1

Bayesian Updating

- Spend the next five minutes drawing similarities between updating in the marbles and urn example and the water on the globe example. Use the appropriate Bayesian terminology.



Conjecture	Ways to produce ●	Previous counts	New count
[○○○○]	0	0	$0 \times 0 = 0$
[●○○○]	1	3	$3 \times 1 = 3$
[●●○○]	2	8	$8 \times 2 = 16$
[●●●○]	3	9	$9 \times 3 = 27$
[●●●●]	4	0	$0 \times 4 = 0$

Deriving the Posterior

- Grid Approximation
- Quadratic Approximation
- Markov Chain Monte Carlo

Classic Example

Imagine you have a test that is able to detect vampirism. The test is able to detect if someone is a vampire 95% of the time. This implies Pr(positive | vampire) = 0.95. One percent of the time, it gets it wrong and says someone is a vampire when they are really mortal (Pr(positive | mortal) = 0.01. We know that vampires make up 0.01% of the population meaning Pr(vampire) = 0.0001.

If you test positive for vampirism, what is the probability that someone is a “bloodsucking immortal”?

Classic Example (Pre Answer?!)

Imagine you have a test that is able to detect vampirism. The test is able to detect if someone is a vampire 95% of the time. This implies Pr(positive | vampire) = 0.95. One percent of the time, it gets it wrong and says someone is a vampire when they are really mortal (Pr(positive | mortal) = 0.01. We know that vampires make up 0.01% of the population meaning Pr(vampire) = 0.0001.

If you test positive for vampirism, what is the probability that someone is a "bloodsucking immortal"?

DISCUSSION QUESTION: What is our power in this situation? What is our Type I error? How does this scenario differ from statistical models we have talked about before?

Classic Example (Pre Answer?!)

Imagine you have a test that is able to detect vampirism. The test is able to detect if someone is a vampire 95% of the time. This implies Pr(positive | vampire) = 0.95. One percent of the time, it gets it wrong and says someone is a vampire when they are really mortal (Pr(positive | mortal) = 0.01. We know that vampires make up 0.01% of the population meaning Pr(vampire) = 0.0001.

$$\Pr(\text{vampire}|\text{positive}) = \frac{\Pr(\text{positive}|\text{vampire})\Pr(\text{vampire})}{\Pr(\text{positive})}$$

$$\begin{aligned}\Pr(\text{positive}) &= \Pr(\text{positive}|\text{vampire})\Pr(\text{vampire}) \\ &\quad + \Pr(\text{positive}|\text{mortal})\left(1 - \Pr(\text{vampire})\right)\end{aligned}$$

//

Classic Example

Imagine you have a test that is able to detect vampirism. The test is able to detect if someone is a vampire **95%** of the time. This implies **Pr(positive | vampire) = 0.95.** One percent of the time, it gets it wrong and says someone is a vampire when they are really mortal (**Pr(positive | mortal) = 0.01.** We know that vampires make up 0.01% of the population meaning **Pr(vampire) = 0.0001**.

$$\Pr(\text{vampire} \mid \text{positive}) = \frac{\Pr(\text{positive} \mid \text{vampire})\,\Pr(\text{vampire})}{\Pr(\text{positive})}$$

LOGICAL!?!?!

0.95 * 0.001 / (0.95 * 0.001 + 0.01(1-0.001)) = ~ 8.7%

Classic Example

Imagine you have a test that is able to detect vampirism. The test is able to detect if someone is a vampire 95% of the time. This implies Pr(positive | vampire) = 0.95. One percent of the time, it gets it wrong and says someone is a vampire when they are really mortal (Pr(positive | mortal) = 0.01. We know that vampires make up 0.01% of the population meaning Pr(vampire) = 0.0001.

1. In a population of 100,000 people, 100 of them are vampires.
2. Of the 100 who ARE vampires, 95 will test positive for vampirism.
3. Of the 99,900 mortals, 999 will test positive for vampirism.

What proportion are actually vampires?

In groups, figure out the P(vampire | positive) with a more simple formulation.

Classic Example

Imagine you have a test that is able to detect vampirism. The test is able to detect if someone is a vampire 95% of the time. This implies Pr(positive | vampire) = 0.95. One percent of the time, it gets it wrong and says someone is a vampire when they are really mortal (Pr(positive | mortal) = 0.01. We know that vampires make up 0.01% of the population meaning Pr(vampire) = 0.0001.

1. In a population of 100,000 people, 100 of them are vampires.
2. Of the 100 who ARE vampires, 95 will test positive for vampirism.
3. Of the 99,900 mortals, 999 will test positive for vampirism.

What proportion are actually vampires?

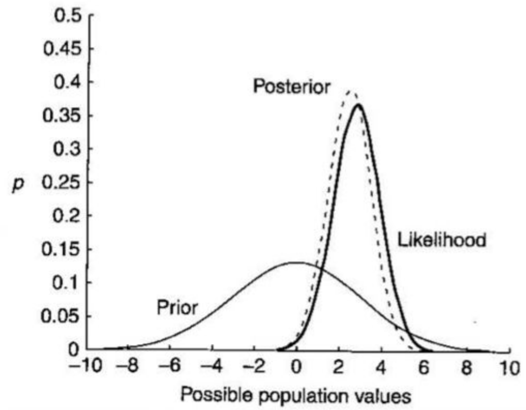
In groups, figure out the P(vampire | positive) with a more simple formulation.

95 / 1094 = ~0.087%

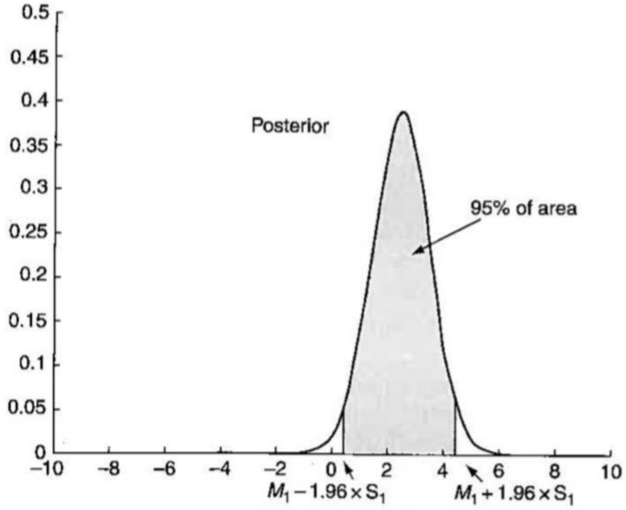
Frequentist (Neyman/Pearson) vs Bayes (Dienes, 2008)

	Meaning Of Probability	Aim	Inference	Long Run Error Rates	Sensitive To
Frequentist Neyman/Pearson	Objective	To provide a reliable decision procedure while controlling long term error rates	Black and white	Controlled	Stopping rules, what is a family of tests, timing of explanation relative to the data
Bayes	Subjective	Indicate how prior probabilities should be changed by data	Continuous degree of posterior	Not guaranteed to be controlled	Prior Opinion

Credibility Interval vs Confidence Interval



The posterior distribution that results from the prior and likelihood we have been considering.



The 95% credibility interval

Discussion Questions

- How does Bayesian statistics differ from Frequentist statistics?
- Define prior, likelihood, and posterior distribution
- Compare and contrast confidence intervals with credible intervals.
- Why should a data scientist care about Bayesian stats?