Naïve Bayesians

Back to Basics Series

13 Feb 2021

Goal

Developing the Bayesian muscle to solve a wide range of problems

Naïve Bayesian Philosophy

Intuitive (Visual) Understanding of the Bayesian Reasoning

Ability to model real world problems in a Bayesian Setting

Starting from Simple Probabilistic modelling

Adapting it in a a Bayesian setting
And moving towards ML models

PyMC3

Fluency in the Calculus of Bayesian Stats & ML model

Season 2: Back to Basics

Episode	Topic
Ep 1	Bayes Theorem
Ep 2, Ep 3	Problems with Binomial Likelihoods
Ep 4	Disease Detection
Ep 5	Naive Bayes Classification
Ep 6	Gaussian Naive Bayes Classification
Ep 7	German Tank Problem
Ep 8	Waiting Times (Continuous Distributions)

Back to Basics

		Canonical Problem	Applications
Ep 1	Bayes Theorem	There are 2 boxes from which cookies can be taken from. Box A and Box B. Box A contains 10 chocolate cookies, Box B contains 5 ginger cookies. Given that you get a chocolate cookie which box was it taken from?	The Shy Librarian Problem Naive Bayes algorithm
Ep 2 Ep 3	Problems with Binomial Likelihoods	You have 2 coins C1 and C2. p(heads for C1) = .7 & P(heads for C2) = 0.6 You flip the coin 10 times. What is the probability that the given coin you picked is C1 given you have 7 heads and 3 tails?	A/B Testing
Ep 4	Disease Detection	A particular disease affects 1% of the population. There is an imperfect test for this disease: The test gives a positive result for 90% of people who have the disease, and 5% of the people who are disease-free. Given a positive test result – what is the probability of having the disease?	COVID Tests (PCR & Antibody)! Fraud Detection
Ep 5	Naive Bayes Classification	Given these words occur in this text what's the probability it's spam?	Any Classification Problem
Ep 6	Gaussian Naive Bayes Classification	Given the weights and heights of basketball players, what’s the probability that person a is a basketball player given weight = w and height = h?	

Back to Basics

		Canonical Problem	Applications
Ep 7	German Tank Problem	Suppose tanks were given a serial number based on the order in which they were manufactured. Given that you've observed a tank with serial number "10", how many tanks were actually manufactured in total?	?
Ep 8	Waiting Times (Continuous Distributions)	Suppose you need to gather 10 patients for a trial. Each signup happens at time t_i (i=1, 10). How long do you have to wait after it took you 3 weeks to accrue 2 signups?	Planning Trials Estimating Queues

Bayes Rule

Posterior Likelihood Prior

$$P(\theta_i \mid D) = \frac{P(D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

Bayes Rule

Posterior　　Likelihood　　Prior

$$P(\theta_i \mid D) = \frac{P(D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

Canonical Problem

Suppose tanks were given a serial number based on the order in which they were manufactured.

You've observed a tank with serial number "41"

How many tanks were manufactured in total?

Canonical Problem Simplified

Suppose N tanks were manufactured.

Each were labelled $d = 1, \ldots, N$ based on the order in which they were manufactured.

You’ve observed a tank with serial number “$d=41$”

What’s N?

Maximum Likelihood Solution

Let's say $N = 100$ tanks were manufactured in total.

What's the probability of observing a tank with the serial number 150?

Maximum Likelihood Solution

Let's say $N = 100$ tanks were manufactured in total.

What's the probability of observing a tank with the serial number 150?

$$P(\ d=150\ )$$

Maximum Likelihood Solution

Let's say $N = 100$ tanks were manufactured in total.

What's the probability of observing a tank with the serial number 150?

$$P(\, d=150 \mid N=100 \,)$$

Maximum Likelihood Solution

Let's say $N = 100$ tanks were manufactured in total.

What's the probability of observing a tank with the serial number 150?

$$P(\, d=150 \mid N=100 \,)$$

$$= 0$$

Maximum Likelihood Solution

Let's say $N = 100$ tanks were manufactured in total.

What's the probability of observing a tank with the serial number 0?

$$P(\, d=0 \mid N=100 \,)$$

$$= 0$$

Maximum Likelihood Solution

Let's say $N = 100$ tanks were manufactured in total.

What's the probability of observing a tank with the serial number -1?

$$P(\, d=-1 \mid N=100 \,)$$

$$= 0$$

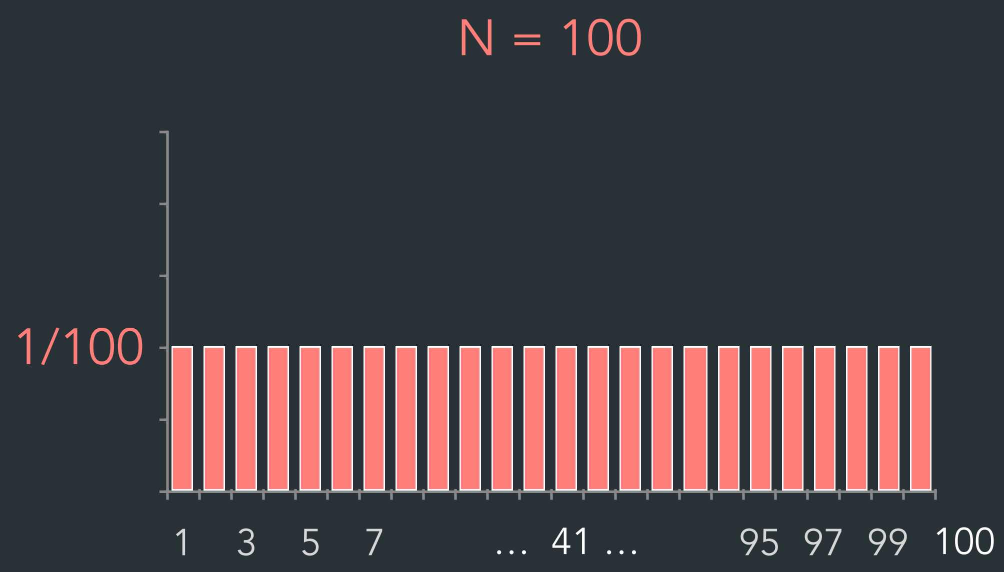
Maximum Likelihood Solution

Let’s say $N = 100$ tanks were manufactured in total.

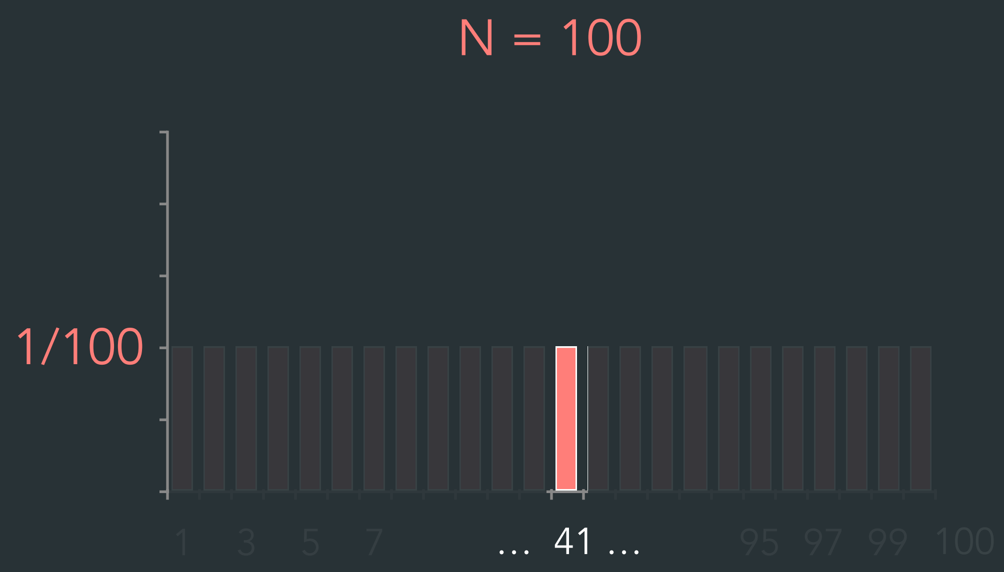
What’s the probability of observing a tank with the serial number 41?

$$P(\, d=41 \mid N=100 \,)$$

Maximum Likelihood Solution

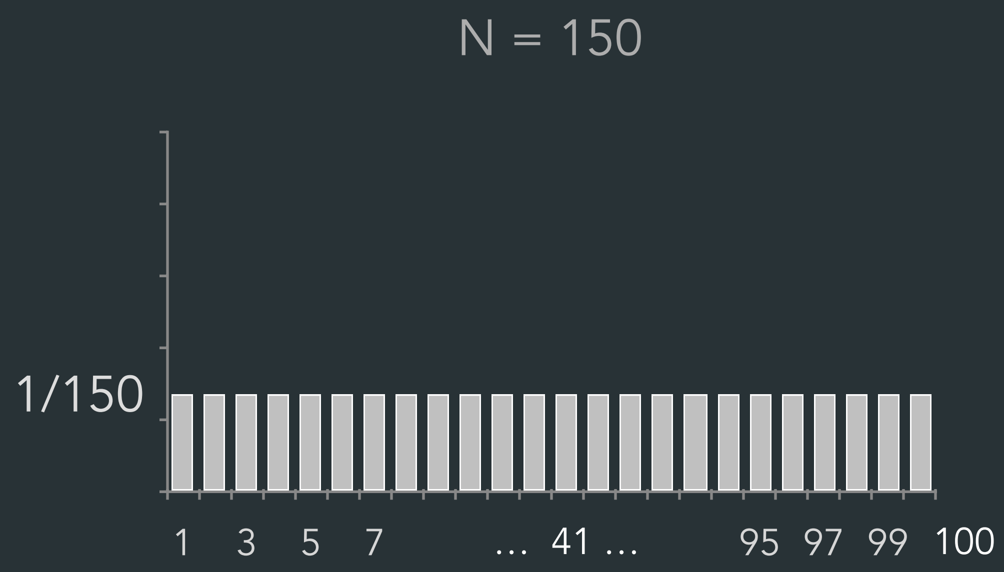


Maximum Likelihood Solution

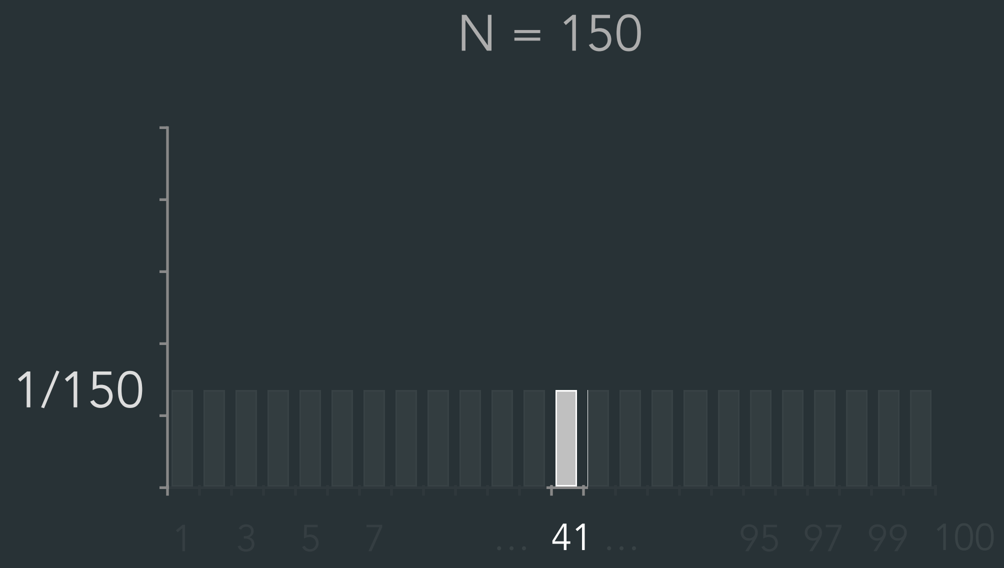


$$P(d{=}41 \mid N{=}100) = \mathbf{1/100}$$

Maximum Likelihood Solution

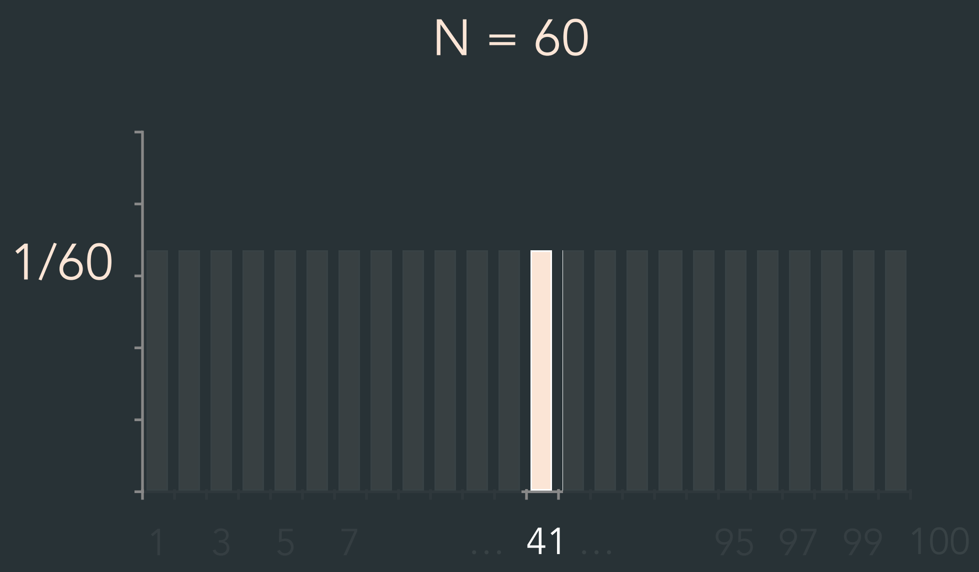


Maximum Likelihood Solution



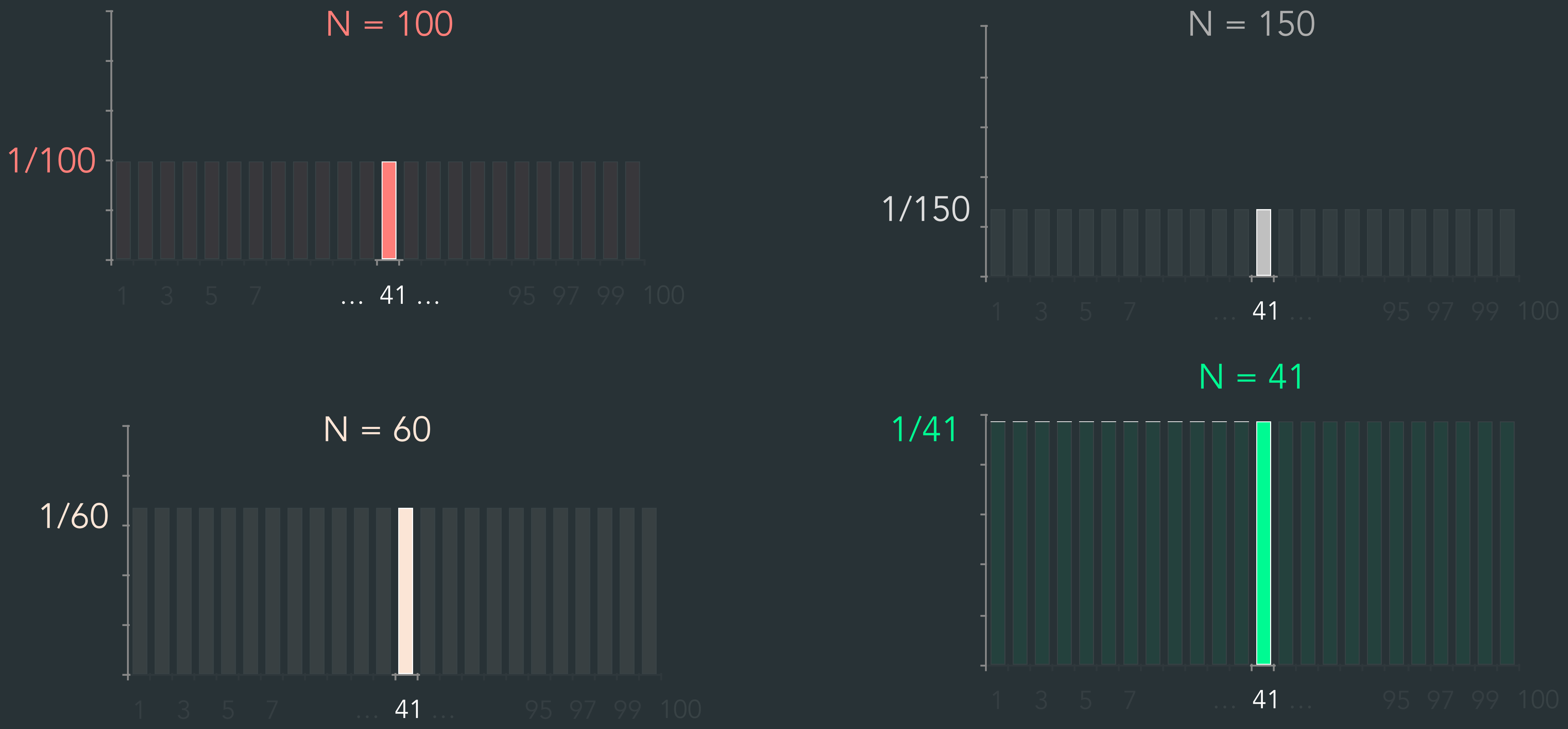
$$P(d{=}41 \mid N{=}100) = \mathbf{1/150}$$

Maximum Likelihood Solution

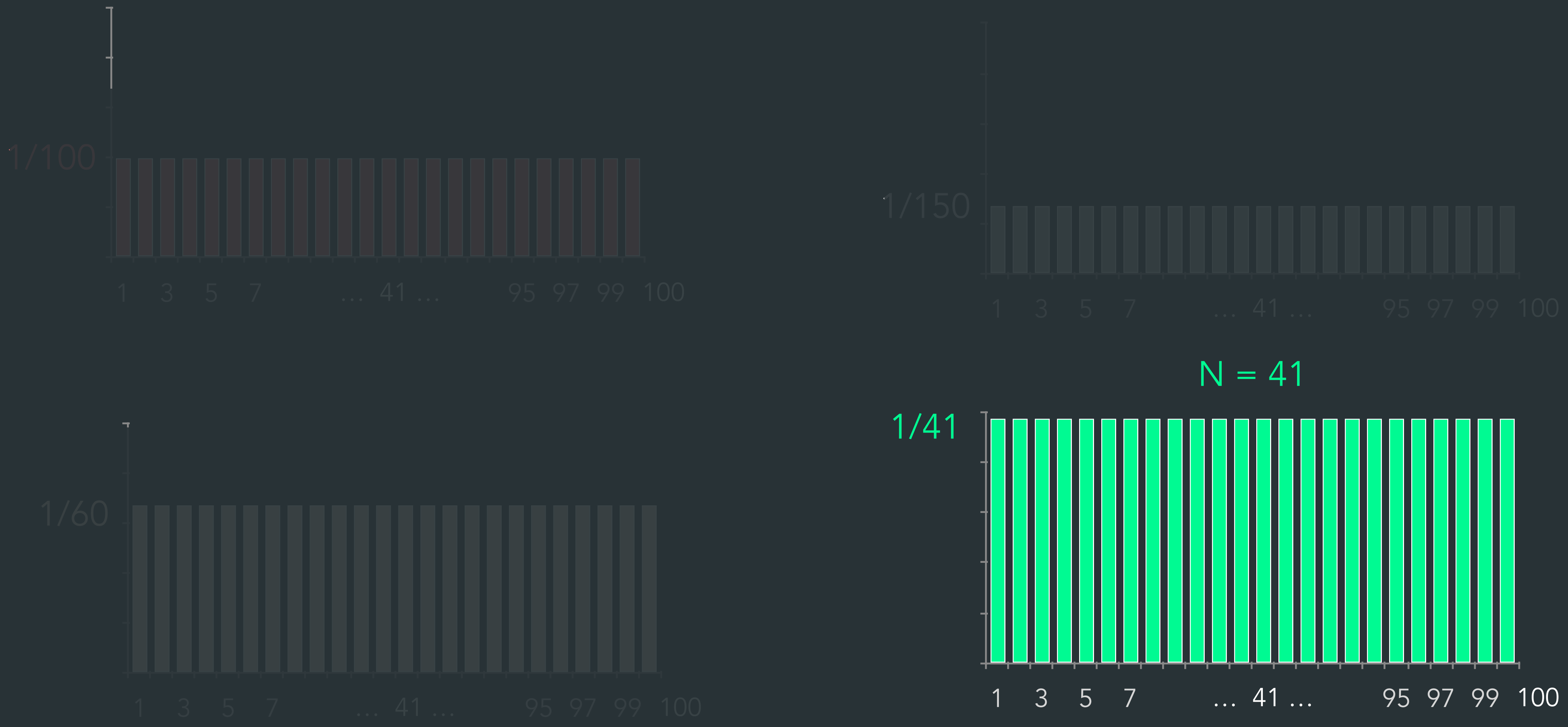


$$P(\, d=41 \mid N\,) = \mathbf{1/N} \quad \text{for } N \geq d$$
$$= 0 \quad \text{otherwise}$$

Maximum Likelihood Solution

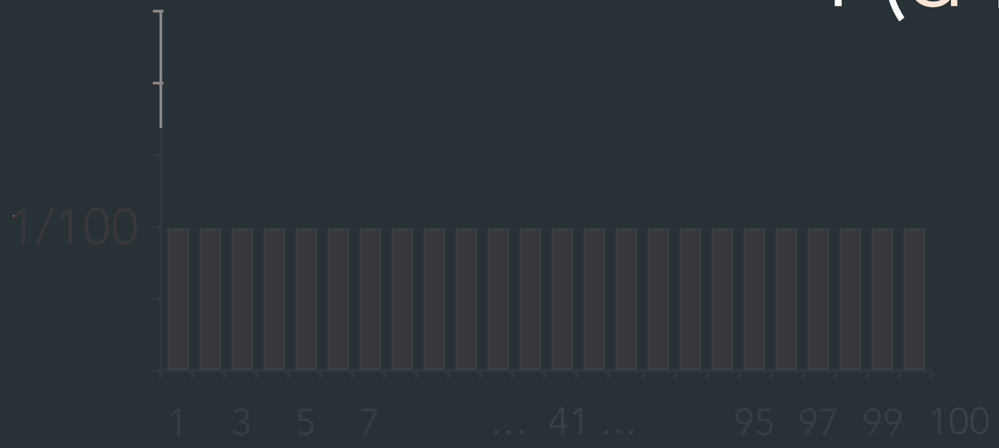


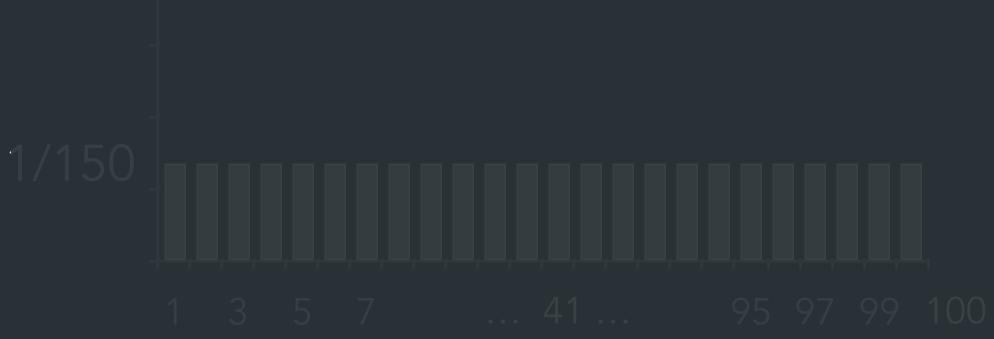
Maximum Likelihood Solution

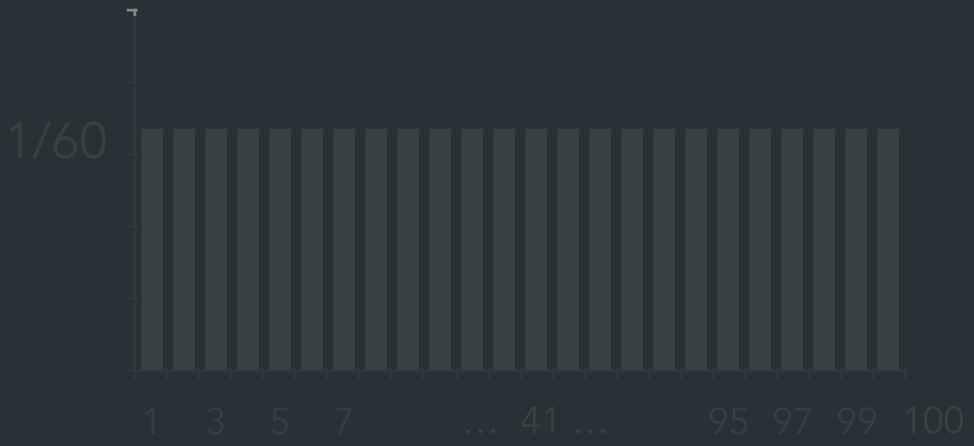


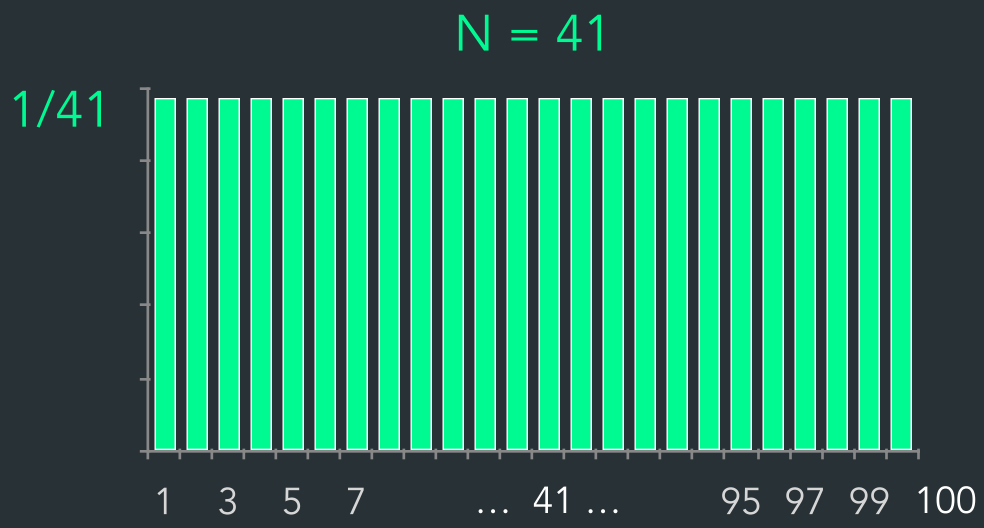
Maximum Likelihood Solution

$$P(d \mid N) = \mathbf{1/N} \quad \text{for } N \geq d$$
$$= 0 \quad \text{otherwise}$$









Maximum Likelihood Solution

$$P(d \mid N) = \mathbf{1/N} \quad \text{for } N \geq d$$
$$= 0 \quad \text{otherwise}$$

MLE : $$\underset{N}{\text{argmax}}\ P(d{=}i \mid N) = 1/N$$

$$P(d{=}i \mid N = i) = 1/i$$

$$N_{MLE} = i$$

Bayes Rule

Posterior Likelihood Prior

$$P(\theta_i \mid D) = \frac{P(D \mid \theta_i)\, P(\theta_i)}{P(D)}$$

Normalising Constant

Bayes Rule

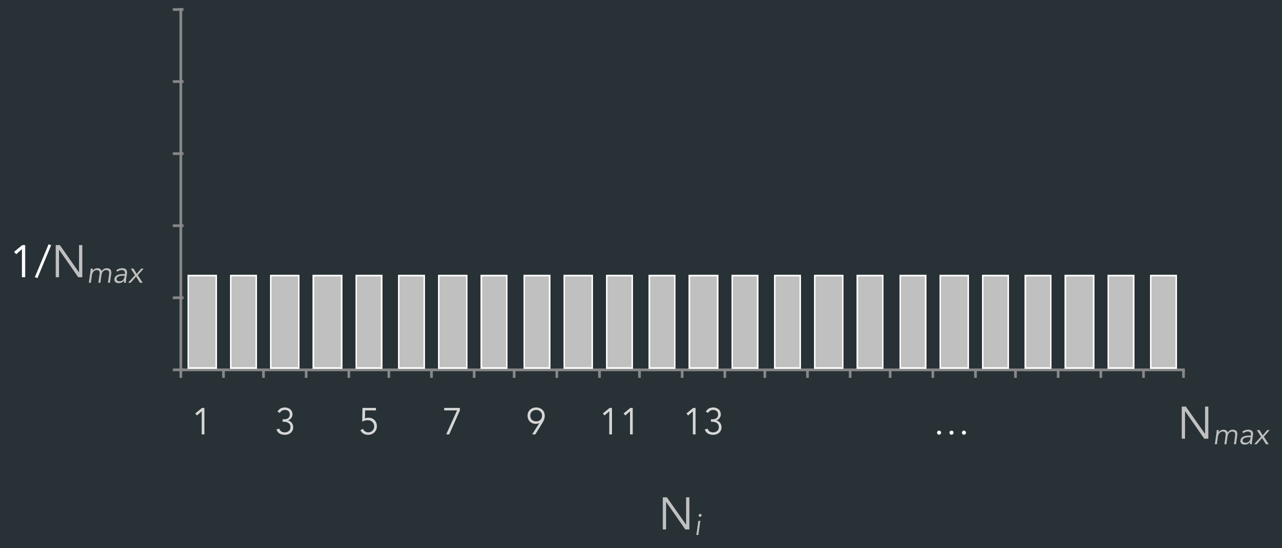
Posterior Likelihood Prior

$$P(N_i \mid d) = \frac{P(d \mid N_i)\, P(N_i)}{P(d)}$$

Normalising Constant

Step 1: Choose a set of *M* hypotheses $N_1, \ldots, N_M$

M is the max number of tanks that one could possibly imagine
e.g. $N_M = 500$

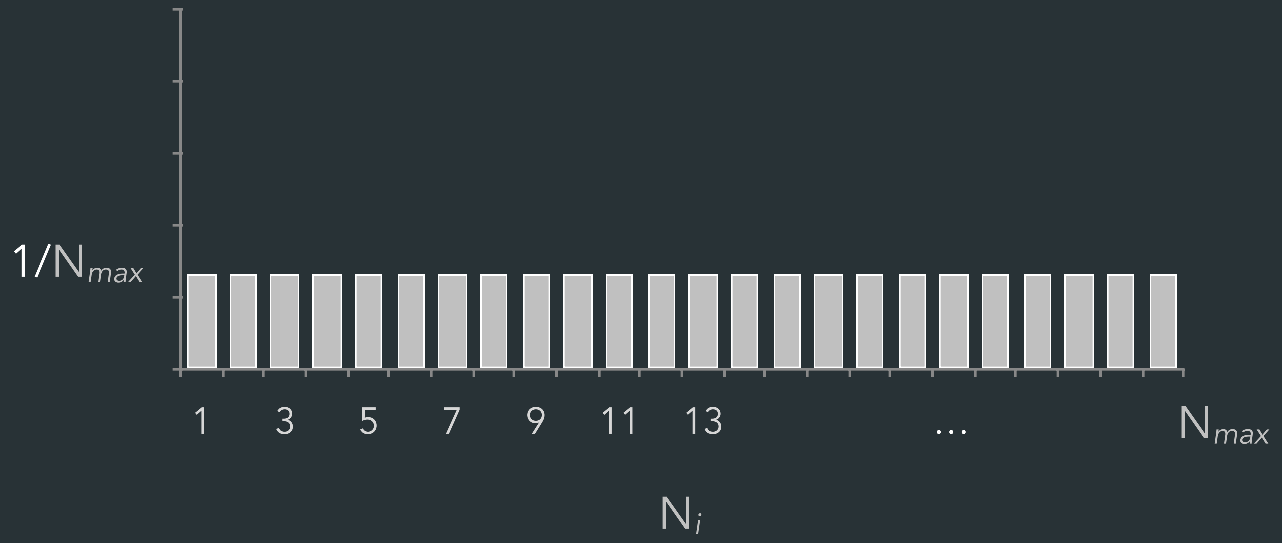


Step 1: Choose a set of *M* hypotheses $N_1, \ldots, N_M$

M is the max number of tanks that one could possibly imagine

e.g. $N_M = 500$

N_i	$P(N_i)$
1	1/500
…	…
40	1/500
41	1/500
42	1/500
43	1/500
…	…
500	1/500
$\sum_{all\ i}$	1



Step 2: Find the likelihood of observing the data for every N_i

$$P(d \mid N_i) = 1/N_i \quad \text{for } N_i \geq d,$$
$$= 0 \quad \text{otherwise.}$$

N_i	$P(N_i)$	$P(d = 41 \mid N_i)$
1	1/500	0
...	...	...
40	1/500	0
41	1/500	1/41
42	1/500	1/42
43	1/500	1/43
...	...	...
500	1/500	1/500
$\sum_{all\ i}$	1	

Step 3: Find the likelihood x prior for every N_i

N_i	P(N_i)	P(d = 41 | N_i)	P(d | N_i) x P(N_i)
1	1/500	0	0
…	…	…	…
40	1/500	0	..
41	1/500	1/41	1/(500x41)
42	1/500	1/42	1/(500x42)
43	1/500	1/43	1/(500x43)
…	…	…	
500	1/500	1/500	1/(500x500)
$\sum_{all\ i}$	1		

Step 4: Find the probability of data P(d)

N_i	$P(N_i)$	$P(d = 41 \mid N_i)$	$P(d \mid N_i) \times P(N_i)$
1	1/500	0	0
…	…	…	…
40	1/500	0	..
41	1/500	1/41	1/(500x41)
42	1/500	1/42	1/(500x42)
43	1/500	1/43	1/(500x43)
…	…	…	
500	1/500	1/500	1/(500x500)
$\sum_{all\ i}$	1		1/(500x41) + … + 1/(500x500)

$$P(d) = \sum_{all\ i} P(d \mid N_i) \times P(N_i)$$

Step 5: Find the posterior probability of data $P(N_i \mid d{=}41)$

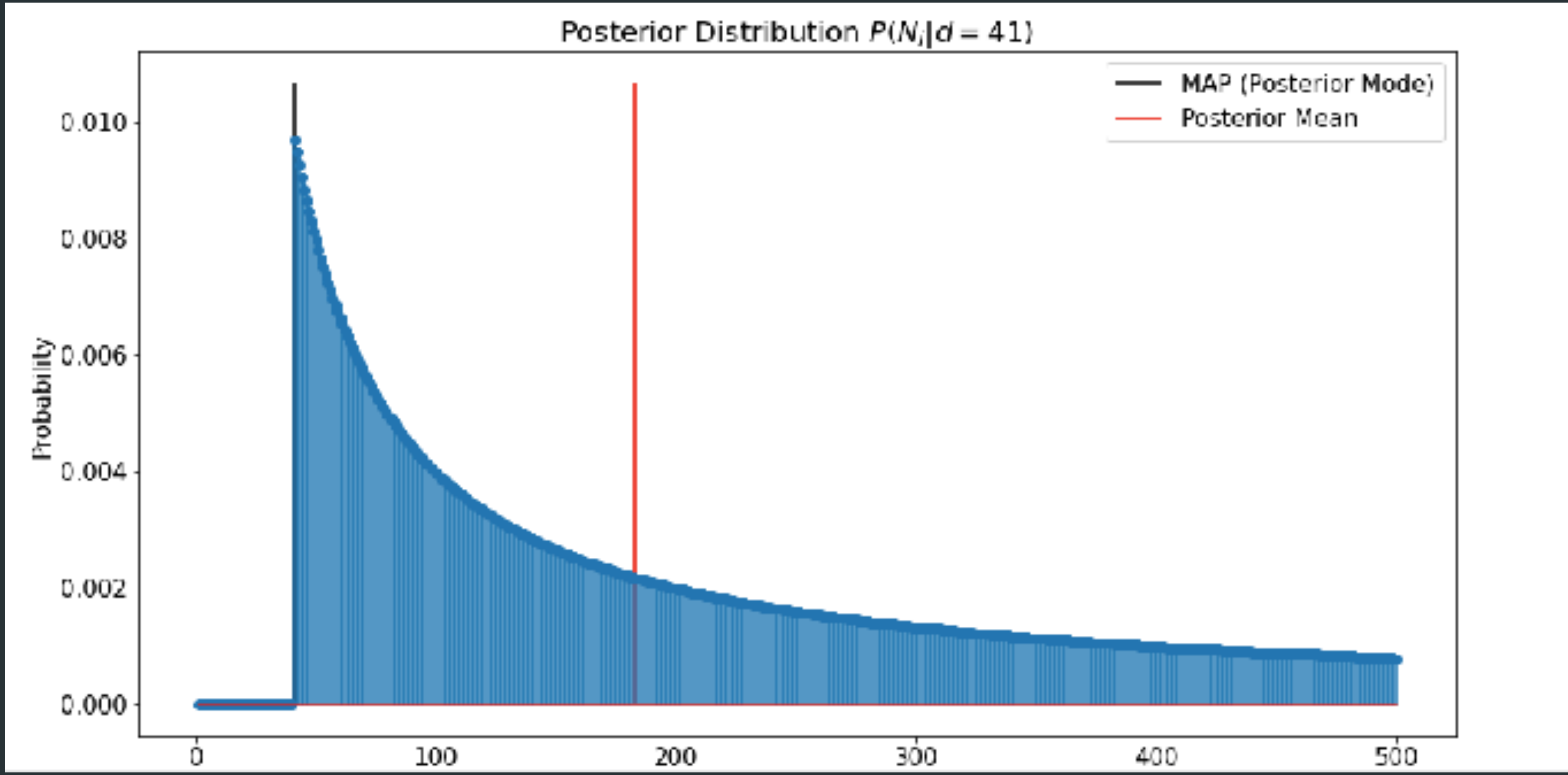
$$P(N_i \mid d{=}41) = \frac{P(d{=}41 \mid N_i)\, P(N_i)}{P(d)}$$

			$\mid N_i) \times P(N_i)$	$P(N_i \mid d = 41)$
			0	0
			…	…
			…	…
			1/(500x41)	[1/(500x41)] / P(d)
			1/(500x42)	[1/(500x42)] / P(d)
			1/(500x43)	[1/(500x43)] / P(d)
…	…	…		
500	1/500	1/500	1/(500x500)	[1/(500x500)] / P(d)
$\sum_{all\ i}$	1		1/(500x41) + … + 1/(500x500)	1

Step 5: Find the posterior probability of data $P(N_i \mid d=41)$

N_i	$P(N_i)$	$P(d = 41 \mid N_i)$	$P(d \mid N_i) \times P(N_i)$	$P(N_i \mid d = 41)$
1	1/500	0	0	0
…	…	…	…	…
40	1/500	0	..	…
41	1/500	1/41	1/(500x41)	[1/(500x41)] / P(d)
42	1/500	1/42	1/(500x42)	[1/(500x42)] / P(d)
43	1/500	1/43	1/(500x43)	[1/(500x43)] / P(d)
…	…	…		
500	1/500	1/500	1/(500x500)	[1/(500x500)] / P(d)
$\sum_{all\ i}$	1		1/(500x41) + … + 1/(500x500)	1

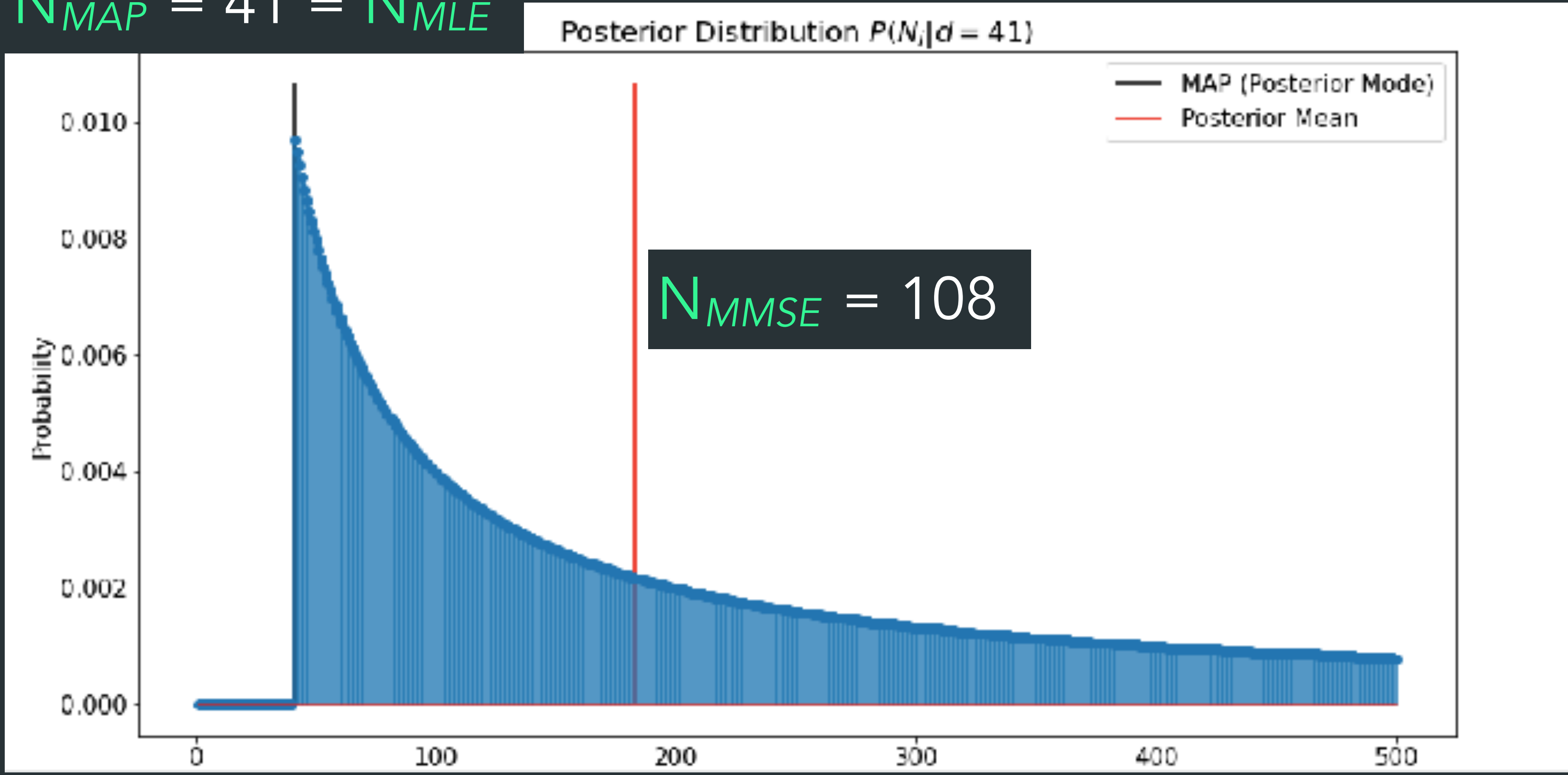
Step 6: Find the point estimate for N



N_i

Step 6: Find the point estimate for N

$N_{MAP} = 41 = N_{MLE}$



$N_{MMSE} = 108$

N_i

Canonical Problem with Multiple Observations

Suppose N tanks were manufactured.

Each were labelled $d = 1, \ldots, N$ based on the order in which they were manufactured.

You've observed 5 tanks with serial numbers "d=41, 31, 25, 39, 32, 37"

What's N?

Canonical Problem with Multiple Observations

Suppose N tanks were manufactured.

Each were labelled $d = 1, \ldots, N$ based on the order in which they were manufactured.

You've observed 2 tanks with serial numbers "d=41, 31"

What's N?

Canonical Problem with Multiple Observations

N_i	$P(N_i \mid d = 41)$
1	0
…	…
40	…
41	[1/(500x41)] / P(d)
42	[1/(500x42)] / P(d)
43	[1/(500x43)] / P(d)
…	
500	[1/(500x500)] / P(d)
$\sum_{all\ i}$	1

Trick:

Use the **posterior** distribution from the d=41 ***as the*** **prior** for the next observation

Canonical Problem with Multiple Observations

Step 2: Find the likelihood of observing the data for every N_i

N_i	$P(N_i \mid d = 41)$	$P(d = 31 \mid N_i)$
1	0	0
…	…	…
40	…	0
41	[1/(500x41)] / P(d)	1/41
42	[1/(500x42)] / P(d)	1/42
43	[1/(500x43)] / P(d)	1/43
…		…
500	[1/(500x500)] / P(d)	1/500
$\sum_{all\ i}$	1	

Canonical Problem with Multiple Observations

Step 3: Find the likelihood x prior for every N_i

Step 4: Find the probability of data P(d)

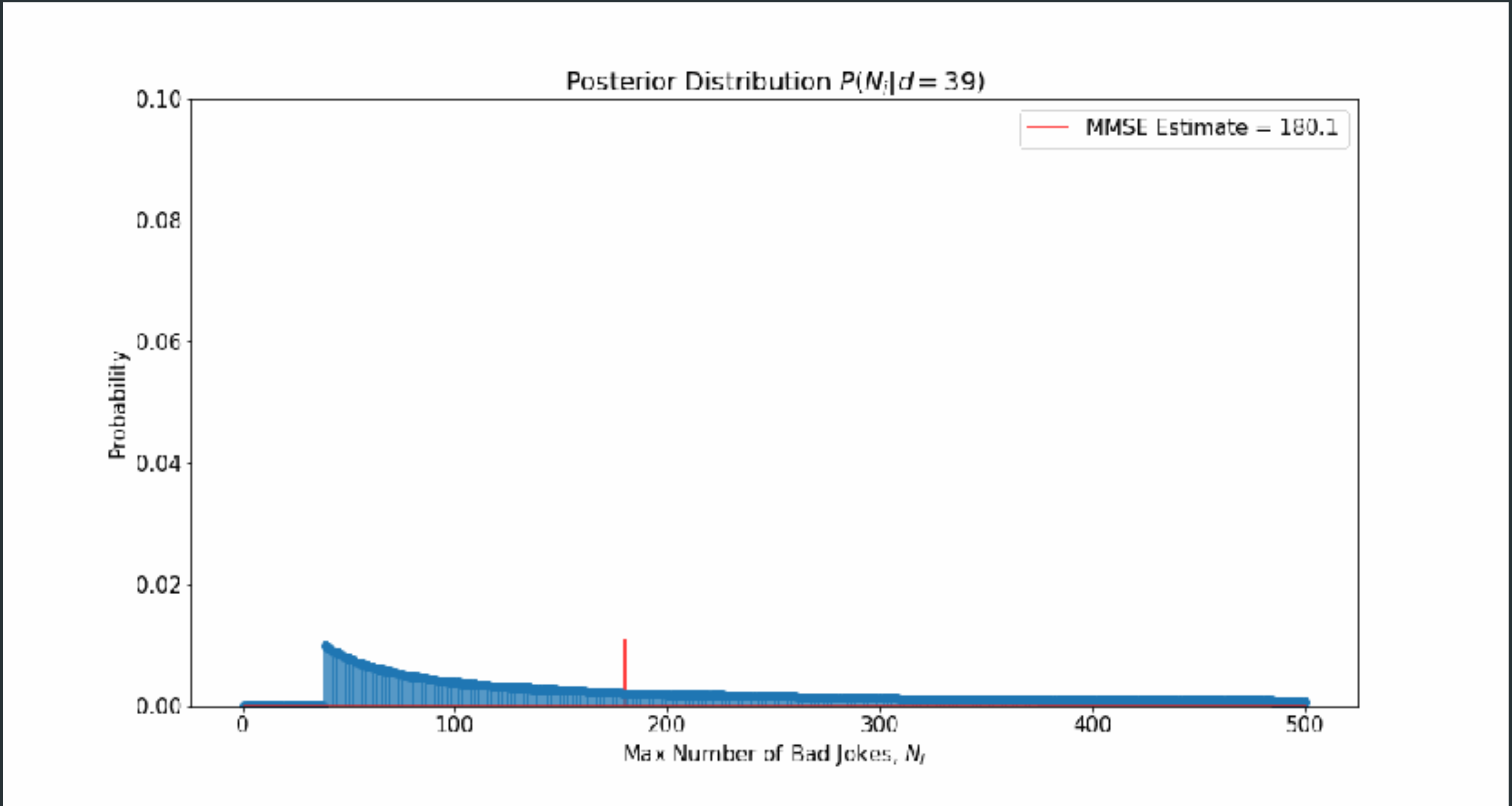
N_i	$P(N_i \mid d = 41)$	$P(d = 31 \mid N_i)$	$P(d \mid N_i) \times P(N_i \mid d = 41)$
1	0	0	
…	…	…	
40	…	0	
41	[1/(500x41)] / P(d)	1/41	
42	[1/(500x42)] / P(d)	1/42	
43	[1/(500x43)] / P(d)	1/43	
…		…	
500	[1/(500x500)] / P(d)	1/500	
$\sum_{all\ i}$	1		P(d)

Canonical Problem with Multiple Observations

Step 5: Find the posterior probability of data $P(N_i \mid d{=}41{,}31)$

N_i	$P(N_i \mid d = 41)$	$P(d = 31 \mid N_i)$	$P(d \mid N_i) \times P(N_i \mid d = 41)$	$P(N_i \mid d = 41, 31)$
1	0	0		
…	…	…		
40	…	0		
41	[1/(500x41)] / P(d)	1/41		
42	[1/(500x42)] / P(d)	1/42		
43	[1/(500x43)] / P(d)	1/43		
…		…		
500	[1/(500x500)] / P(d)	1/500		
$\sum_{all\ i}$	1		P(d)	1

Canonical Problem with Multiple Observations



References

Think Stats: Probability and Statistics for Programmers (Allen B. Downey)
https://greenteapress.com/thinkstats/html/thinkstats009.html#toc75