A CURSORY GLANCE AT BAYESIAN ANALYSIS

SPECIAL TOPICS IN A.I. AND DATA SCIENCE

"Classical data analysts need a large bag of clever tricks to unleash on their data, but Bayesians only ever really need one."

(author unknown)

OUTLINE

Part 1

1. Plausible Reasoning
2. The Rules of Probability
3. Bayes' Theorem
4. Example: the Fair (?) Coin
5. Example: the Salary Question
6. Example: Money ($ Bill Y'All)

Part II

7. Marginalization
8. Prior Distributions
9. Model Selection
10. Naïve Bayes Classification
11. Bayesian Inference
12. MCMC and Numerical Methods

PLAUSIBLE REASONING

A CURSORY GLANCE AT BAYESIAN ANALYSIS

> “A decision was wise, even though it lead to disastrous consequences, if the evidence at hand indicated it was the best one to make; and a decision was foolish, even though it lead to the happiest possible consequences, if it was unreasonable to expect those consequences.”
>
> Herodotus, in Antiquity

PLAUSIBLE REASONING

Consider the following scenario [Jaynes, 2003]:

- you are walking down a deserted street at night
- you hear a security alarm, look across the street, and see a store with a broken window
- someone wearing a mask crawls out of the broken window with a bag full of smart phones

You conclude that the person crawling out of the store is stealing merchandise from the store.

PLAUSIBLE REASONING

How do you come to that conclusion? It **cannot** come from a logical deduction based on evidence.

Indeed, the person crawling out of the store **could** have been its owner who, upon returning from a costume party, realized that they had misplaced their keys just as a passing truck was throwing a brick in the store window, triggering the security alarm. The owner then went into the store to retrieve items before they could be stolen, which is when you happened unto the scene.

The original reasoning process is not **deductive**, but it is at least **plausible**.

DEDUCTIVE VS. PLAUSIBLE REASONING

Deductive (ideal) reasoning:

If A is true, then B is true
A is true

???

If A is true, then B is true
B is false

???

Inductive (plausible) reasoning:

If A is true, then B is true
B is true

???

If A is true, then B is true
A is false

???

DEDUCTIVE VS. PLAUSIBLE REASONING

Deductive (ideal) reasoning:

If A is true, then B is true
A is true

B is true

If A is true, then B is true
B is false

A is false

Inductive (plausible) reasoning:

If A is true, then B is true
B is true

A is more plausible (why?)

If A is true, then B is true
A is false

B is less plausible (why?)

DEDUCTIVE VS. PLAUSIBLE REASONING

Inductive (plausible) reasoning:

If A is true, then B is more plausible
B is true

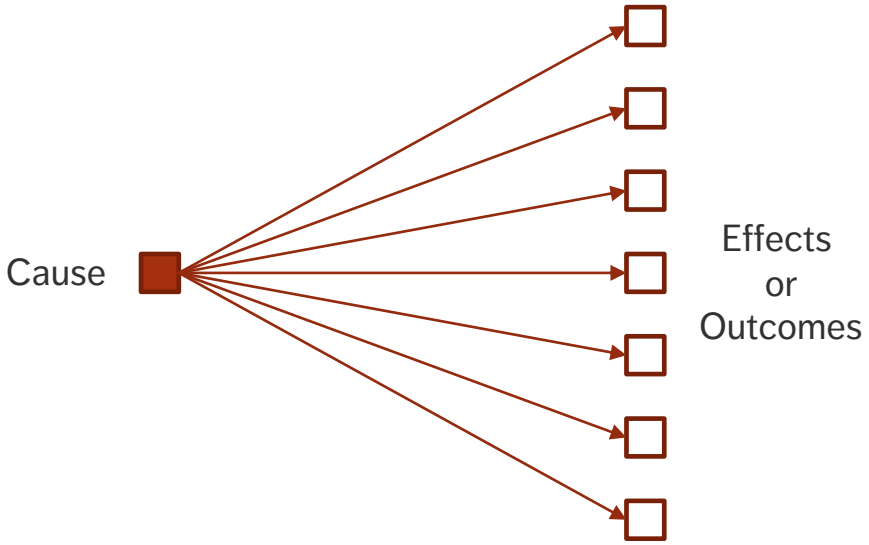
A is more plausible

If “the person is a thief” (A is true), you would not be surprised to “see them crawling out of the store with a bag of phones” (B is plausible). You do “see them crawling out of the store with a bag of phones” (B is true). Therefore, you would not be surprised if “the person were a thief” (A is plausible).

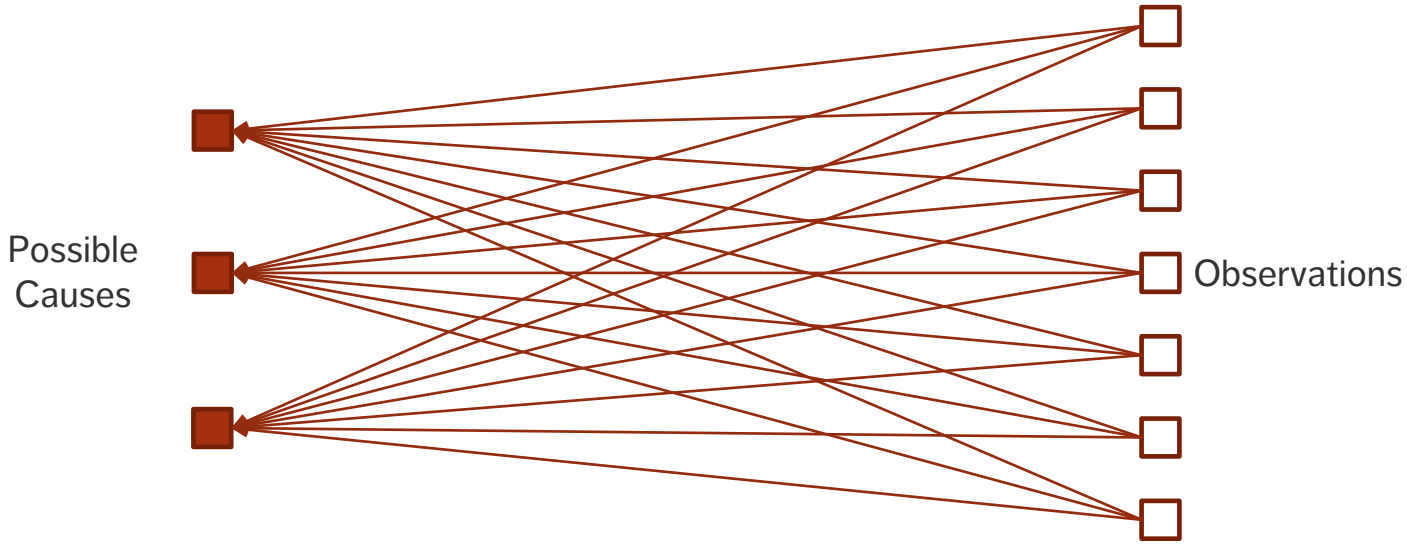
[Sivia, Skilling, *Data Analysis: A Bayesian Tutorial*]

DEDUCTIVE VS. PLAUSIBLE REASONING

Deductive reasoning



Plausible reasoning



DISCUSSION

In Tom Stoppard's 1966 play *Rosencrantz and Guildenstern are Dead*, the main characters bet on coin flips. Rosencrantz wins by flipping heads 92 times in a row.

This result is of course not impossible, but is it plausible? If this happened to you, what would you conclude?

THE RULES OF PROBABILITY

A CURSORY GLANCE AT BAYESIAN DATA ANALYSIS

WHAT IS PROBABILITY?

Inductive reasoning requires methods to evaluate the validity of various propositions.

For Bernouilli, Bayes, and Laplace (1700's to 1800's), a proposition's probability represents the **degree-of-belief** in the proposition (i.e., its plausibility).

Subsequent scholars found this vague and subjective (how can you be sure that my degree-of-belief matches yours?) and they redefined probability as the **long-run relative frequency** of an event, given infinite repeated trials.

WHAT IS PROBABILITY?

A forecast calling for rain with 90% probability doesn't mean the same thing to **Bayesians** and **frequentists**:

- in the Bayesian framework, this means that the forecaster is 90% certain that it will rain;
- in the frequentist framework, this means that, historically, it rained in 90% of the cases when the conditions were as they currently are.

The Bayesians framework is more aligned with how humans understand probabilities (92 heads in a row probably mean that that the coin is biased, right?), but how can we be certain that the **degree-of-belief** is a well-defined concept?

WHAT IS PROBABILITY?

As it happens, there is a well-defined way to determine the rules of probability, based on a small list of axioms [Jaynes, 2003; Cox, 1946]:

1. if a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result;
2. all (known) evidence relevant to a question must be taken into consideration;
3. equivalent states of knowledge must be assigned the same probabilities;
4. if we specify how much we believe something is true, we have implicitly specified how much we believe it's false, and
5. if we have specified our degree-of-belief in a first proposition, and then our degree-of-belief in a second proposition if we assume the first one is true, then we have implicitly specified our simultaneous degree-of-belief in both propositions being true.

RULES OF PROBABILITY

Let I denote relevant background information; X, Y, Y_k denote various propositions, and $-X$ denote the proposition that X is false.

The **plausibility** of X given I is denoted by $P(X|I)$, ranging from 0 (false) to 1 (true).

Sum Rule: $P(X|I) + P(-X|I) = 1$

Product Rule: $P(X, Y|I) = P(X|Y, I) \times P(Y|I)$

Bayes' Theorem: $P(X|Y, I) \times P(Y|I) = P(Y|X, I) \times P(X|I)$

Marginalization Rule: $P(X|I) = \sum P(X, Y_k|I)$, where $\{Y_k\}$ are exhaustive, disjoint

CONDITIONAL PROBABILITIES

We then have an interest in determining the likelihood of an event occurring **given that another event** (or series of events) **has occurred.**

Examples include:

- the probability that a train arrives on time given that it left on time
- the probability that a PC crashes given the operating system installed
- the probability that a bit is transmitted over a channel is received as a 1 given that the bit transmitted was a 1
- the probability that a website is visited given its number of in-links
- the rules of probability!

CONDITIONAL PROBABILITIES

A **conditional probability** is the probability of an event taking place given that another event occurred.

The conditional probability of A given B, $P(A|B)$, is defined as

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

The probability that two events A and B both occur simultaneously is obtained by applying the multiplication rule:

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

EXERCISE – CONDITIONAL PROBABILITIES

Example (a classic): a family has two children (not twins). What is the probability that the youngest child is a girl given that at least one of the children is a girl? Assume that boys and girls are equally likely to be born.

EXERCISE – CONDITIONAL PROBABILITIES

We will first try to answer this question by generating a number of trials and identifying successful events (a frequentist approach?)

EXERCISE – CONDITIONAL PROBABILITIES

Example (a classic): a family has two children (not twins). What is the probability that the youngest child is a girl given that at least one of the children is a girl? Assume that boys and girls are equally likely to be born.

Solution: Let A and B be the events that the youngest child is a girl and that at least one child is a girl, respectively:

$$A = \{GG, BG\}, \qquad B = \{GG, BG, GB\}$$

Then $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{3}$ (not ½, as one might naively assume).

BAYES' THEOREM

A CURSORY GLANCE AT BAYESIAN DATA ANALYSIS

BAYES' THEOREM

The sum rule and the product rules are the **basic rules of probability**.

Bayes' Theorem and the **Marginalization Rule** are simple corollaries of these basic rules.

Bayes' Theorem is sometimes written is a slightly different form

$$P(X|Y,I) = \frac{P(Y|X,I) \times P(X|I)}{P(Y|I)}$$

BAYES' THEOREM

Set-up: assume that an experiment has been conducted to determine the degree of validity of a particular hypothesis, and that experimental data has been collected.

The central data analysis question: given everything that was known *prior* to the experiment, does the collected data support (or invalidate) the hypothesis?

Throughout, let X denote the proposition that the hypothesis in question is true, let Y denote the proposition that the experiment yielded the actual observed data, let I denote (as always) the relevant background information.

BAYES' THEOREM

Central data analysis question (reprise):

What is the value of $P(\text{hypothesis is true} \mid \text{observed data}, I)$?

Problem: this is nearly always impossible to compute directly.

Solution: using Bayes' Theorem,

$$P(\text{hypothesis} \mid \text{data}, I) = \frac{P(\text{data} \mid \text{hypothesis}, I) \times P(\text{hypothesis} \mid I)}{P(\text{data} \mid I)},$$

it may be that the terms on the right are easier to compute.

[Silver, *The Signal and the Noise*, pp. 247-248]

EXERCISE – WORLD TRADE CENTER

"Consider a somber example: the September 11 attacks. Most of us would have assigned almost no probability to terrorists crashing planes into buildings in Manhattan when we woke up that morning. But we recognized that a terror attack was an obvious possibility once the first hit the World Trade Center. And we had no doubt we were being attacked once the second tower was hit. Bayes' Theorem can replicate this result."

[Silver, *The Signal and the Noise*, pp. 247-248]

EXERCISE – WORLD TRADE CENTER (1ST PLANE)

PRIOR PROBABILITY		
Initial estimate of how likely it is that terrorists would crash planes into Manhattan skyscrapers	$P(B\|I) = x$	0.005%
A NEW EVENT OCCURS: FIRST PLANE HITS WTC		
Probability of plane hitting if terrorists are attacking Manhattan skyscrapers	$P(A\|B, I) = y$	95%+
Probability of plane hitting if terrorists are *not* attacking Manhattan skyscrapers (i.e. accident)	$P(A\|\bar{B}, I) = w$	0.008%*
POSTERIOR PROBABILITY		
Revised estimate of probability of terror attack, given first plane hitting WTC	$P(B\|A, I) = \frac{yx}{yx + w\,(1 - x)}$	37%+

*2 incidents in the previous 25,000 days

[Silver, *The Signal and the Noise*, pp. 247-248]

EXERCISE – WORLD TRADE CENTER (2ND PLANE)

PRIOR PROBABILITY		
Revised estimate of probability of terror attack (now that we know about the first plane hitting WTC)	$P(B\|I) = x^{\#}$	37%+
A NEW EVENT OCCURS: FIRST PLANE HITS WTC		
Probability of plane hitting if terrorists are attacking Manhattan skyscrapers	$P(A\|B,I) = y$	95%+
Probability of plane hitting if terrorists are *not* attacking Manhattan skyscrapers (i.e. accident)	$P(A\|\bar{B},I) = w$	0.008%*
POSTERIOR PROBABILITY		
Revised estimate of probability of terror attack, given second plane hitting WTC	$P(B\|A,I) = \frac{yx^{\#}}{yx^{\#} + w\,(1 - x^{\#})}$	99.99%+

*2 incidents in the previous 25,000 days

BAYES' THEOREM

In the vernacular, the probabilities

- $P(\text{hypothesis} \mid I)$ of the hypothesis being true prior to the experiment is the **prior**;
- $P(\text{hypothesis} \mid \text{data}, I)$ of the hypothesis being true once the experimental data is taken into account is the **posterior**;
- $P(\text{data} \mid \text{hypothesis}, I)$ of the experimental data being observed assuming that the hypothesis is true is the **likelihood**, and
- $P(\text{data} \mid I)$ of the experimental data being observed independently of any hypothesis is the **evidence**.

A given hypothesis includes a (potentially implicit) model which can be used to compute or approximate the **likelihood**.

BAYES' THEOREM

Determining the **prior** is a source of considerable controversy

- conservative estimates (uninformative priors) often lead to reasonable results
- in the absence of information, go with maximum entropy prior

The **evidence** is harder to compute on theoretical grounds – evaluating the probability of observing data requires access to some model as part of I. Either

- that model was good, so there's no need for a new hypothesis
- that model was bad, so we dare not trust our computation

BAYES' THEOREM

Thankfully, the evidence is rarely required on problems of parameter estimation (although it is crucial for model selection):

- prior to the experiment, there are numerous competing hypotheses
- the priors and likelihoods will differ, but not the evidence
- the evidence is not needed to differentiate the various hypotheses

Bayes' Theorem is often presented as

$$P(\text{hypothesis} \mid \text{data}, I) \propto P(\text{data} \mid \text{hypothesis}, I) \times P(\text{hypothesis} \mid I)$$

or simply as posterior $\propto$ likelihood $\times$ prior, that is to say, **beliefs should be updated in the presence of new information**.

DISCUSSION

What would it take for you to update ...

- ... your belief in the existence/non-existence of a deity?
- ... your belief in the shape of the Earth?
- ... your political affiliation?
- ... your allegiance to a sport team?
- ... your belief in the effectiveness of homeopathic remedies?
- ... your belief in the effectiveness of Bayesian analysis?

EXERCISE – FALSE POSITIVE TESTING

Suppose that a test for a particular disease has a very high success rate. If a patient

- has the disease, the test accurately reports a 'positive' with probability 0.99;
- does not have the disease, the test accurately reports a 'negative' with probability 0.95.

Assume further that only 0.1% of the population has the disease. What is the probability that a patient who tests positive does not in fact have the disease?

EXERCISE – FALSE POSITIVE TESTING

Solution: let D be the event that the patient has the disease, and A be the event that the test is positive. According to Bayes' Theorem, the probability of a **true positive** is

$$P(D|A,I) = \frac{P(A|D,I)\times P(D|I)}{P(A|I)} = \frac{P(A|D,I)\times P(D|I)}{P(A|D,I)\times P(D|I) + P(A|-D,I)\times P(-D|I)}$$
$$= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} \approx 0.019;$$

the probability of a **false positive** is thus $1 - 0.019 \approx 0.981$.

Despite the apparent high accuracy of the test, the incidence of the disease is so low (1 in a 1000) that the vast majority of patients who test positive (98 in 100) do not have the disease (20 times the proportion before the outcome of the test is known).

EXERCISE – DRIVING CONDITIONS

A road safety analyst has access to a dataset of fatal vehicle collisions (such as the NCDB) on roads in a specific region.

The dataset is built using police reports, and it contains relevant collision information such as:

- the severity of the collision, the age of the drivers, the number of passengers in each vehicle, the date and time of the collision, weather and road conditions, blood alcohol content (BAC), etc.

Let us further assume that the analyst has access to aggregated weather data and R.I.D.E. (sobriety checkpoint) reports for that region.

EXERCISE – DRIVING CONDITIONS

Some information may be missing from the police reports at a given moment (perhaps the coroner has not yet had the chance to determine the BAC level, or some of the data may have been mistakenly erased and/or corrupted).

For some collisions, we may need to answer either or both of the following questions:

- did alcohol play a role in the collision?
- did "bad" weather play a role in the collision?

As usual, let I denote all relevant information relating to the situation, such as the snowy months of the year, the incidence of impaired driving in that region, etc.

EXERCISE – DRIVING CONDITIONS

Our analyst will consider 3 propositions:

- A: a fatal collision has occurred
- B: weather and road conditions were bad
- C: the BAC level of one of the drivers was above 0.08% per volume

and may have an interest in $P(B|A,I)$, $P(C|A,I)$, $P(B,C|A,I)$, $P(B,-C|A,I)$, or $P(-B,C|A,I)$, for instance.

EXERCISE – DRIVING CONDITIONS

1. Derive an expression to compute the probability that “bad” weather and road conditions were present at the time of the collision.

2. **A Mild Winter** scenario: during a mild winter, “bad” weather affected regional road conditions 5% of the time. The analyst knows from other sources that the probabilities of fatal collisions given “bad” and “good” weather conditions in the region over the winter are 0.01% and 0.002%, respectively. If a fatal collision occurred on a regional road that winter, what is the probability that the weather conditions were “bad” on that road at that time? Is the result surprising?

EXERCISE – DRIVING CONDITIONS

3. **Not quite as Mild a Winter** scenario: let's assume that the winter was not quite as mild (perhaps "bad" weather affected regional road conditions 10% of the time, say). If a fatal collision occurred on a regional road that winter, what is the probability that the weather conditions were "bad" on that road at that time? How much of a jump are you expecting compared to question 2?

4. Just how rough of a winter would be necessary before we conclude that a given fatal collision was more likely to have occurred in "bad" weather?

EXERCISE – DRIVING CONDITIONS

5. In what follows, we assume that the analyst does not have access to other sources from which to derive the individual probabilities of fatal collisions given "bad" and "good" weather conditions in the region. Instead, the analyst has access to data that suggests that the probability of a fatal collision in "bad" weather is k times as high as the probability of a fatal collision in "good" weather. Let the probability of "bad" weather be $w \in (0,1)$. Derive an expression for the probability that the weather conditions were "bad" on that road at that time, given that a fatal collision occurred, in terms of k and w.

EXERCISE – DRIVING CONDITIONS

6. **Really Rough Winter** scenario: during a really rough winter, "bad" weather affected road conditions with probability $w = 0.2$. Determine the probabilities that there were "bad" weather conditions given a fatal collision under 4 different values: $k = 0.1, k = 1, k = 10, k = 100$. Which of these scenarios is most likely?

7. In the next scenario, we assume that the traffic flow changes depending on the weather; while some individuals need to be on the roads no matter the conditions, others might tend to avoid the roads when the conditions are "bad". Make whatever assumptions are necessary and analyze the situation as you have done in the previous questions.

8. Repeat the process for the other conditional probabilities of interest.

EXERCISE – MONTY HALL PROBLEM

(Another classic) A lifetime's supply of poutine is placed randomly behind one of three identical doors. You are asked to pick a door. One of the doors you have not selected is opened, revealing an empty room. You are given the option of changing your pick. What is your optimal strategy?

1. Determine the ideal strategy using a simulation.
2. Analyze a similar situation (for 100 doors instead of 3) using Bayes' Theorem.
3. Analyze the situation using Bayes' Theorem.

EXAMPLE: THE FAIR (?) COIN

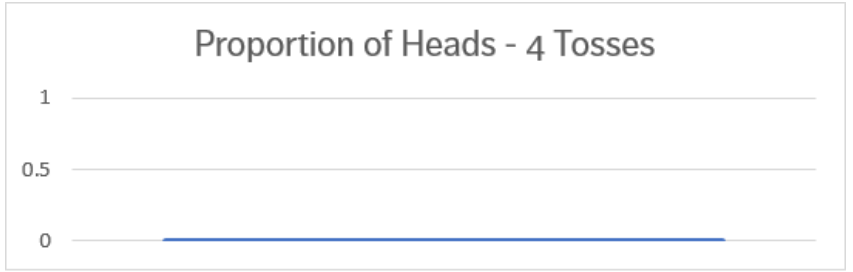
A CURSORY GLANCE AT BAYESIAN DATA ANALYSIS

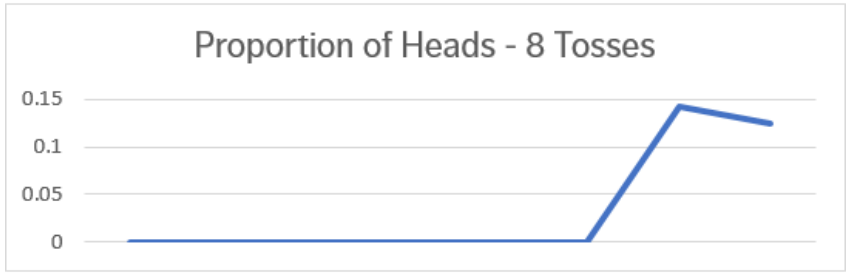
THE FAIR (?) COIN – SET-UP

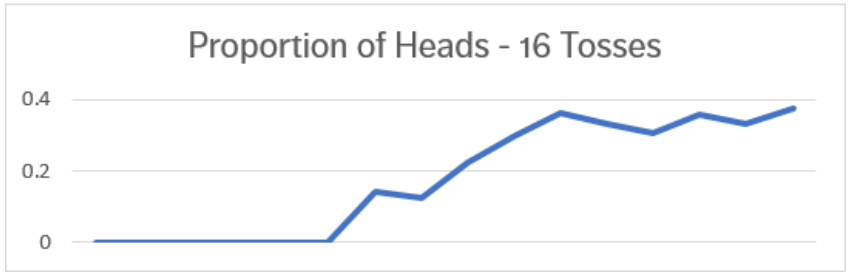
I brought back a souvenir coin from a trip to a strange and distant land.

I have been flipping it pretty much non-stop since I've returned. You can see the proportion of heads I obtained for 4, 8, and 16 tosses.

At first, I thought the coin might be biased, but the proportion of heads seems to inch its way towards 50%...





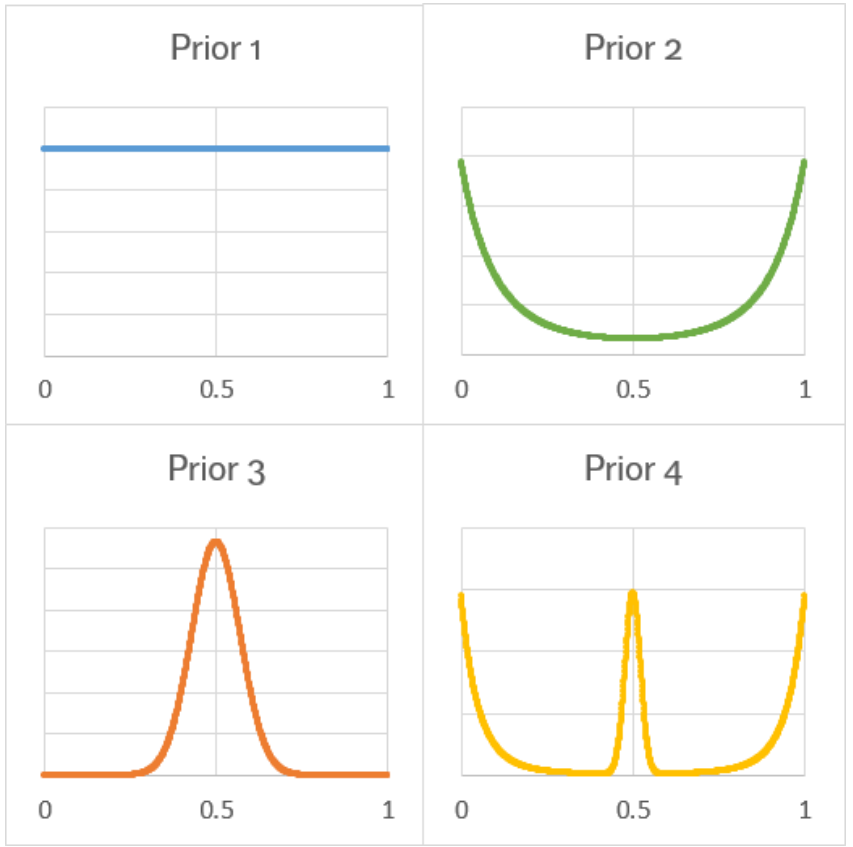


THE FAIR (?) COIN – PRIORS

Perhaps the coin isn't fair, coming as it does from a strange and distant land...

Let's denote the coin's **bias** by H, i.e. the probability of flipping a head on a toss ($H \approx 0.5$: regular unbiased coins, $H \approx 0$ or $H \approx 1$: highly biased coins).

A **prior** for this scenario is a function $P(\text{bias} = H) = P(H|I)$, for $0 \leq H \leq 1$.

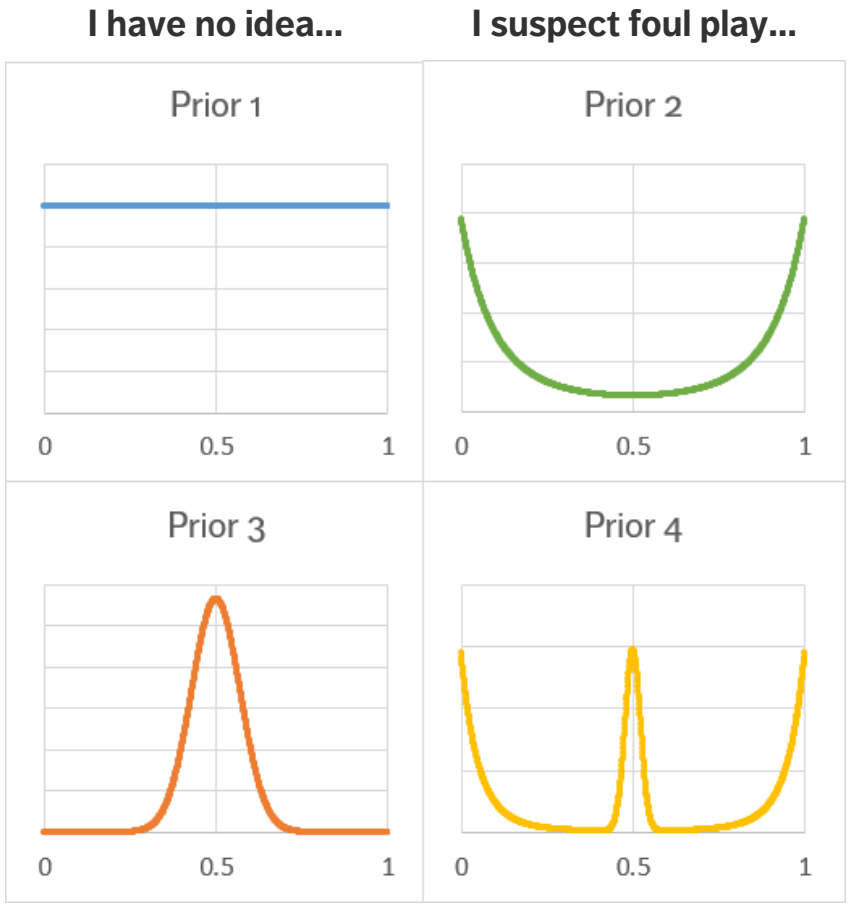


THE FAIR (?) COIN – PRIORS

Perhaps the coin isn't fair, coming as it does from a strange and distant land...

Let's denote the coin's **bias** by H, i.e. the probability of flipping a head on a toss ($H \approx 0.5$: regular unbiased coins, $H \approx 0$ or $H \approx 1$: highly biased coins).

A **prior** for this scenario is a pdf function $P(\text{bias} = H) = P(H|I)$, for $0 \leq H \leq 1$.



THE FAIR (?) COIN – PRIORS

Why are we working with functions for the prior? In the previous example (Sept. 11 attacks), we only provided a number $P(B|I) = 0.005\%$.

In fact, we had provided a (discrete) function:

$$P(B = x|I) = \begin{cases} 00.005\% & \text{if } x = \text{TRUE} \\ 99.995\% & \text{if } x = \text{FALSE} \end{cases}$$

THE FAIR (?) COIN – LIKELIHOOD

Let's assume that the coin has been tossed N times in total, and that K heads have been recorded. In this scenario, Bayes' Theorem takes the form:

$$P\left(\text{bias} = H \,\middle|\, K \text{ heads}, N \text{ tosses}; I\right) \propto P(K \text{ heads}, N \text{ tosses}|\ \text{bias} = H, I) \times P(\text{bias} = H|I).$$

The **likelihood** is the probability of observing K heads in N tosses with a bias of H. If, as part of I, the tosses are independent (i.e. the result of one toss does not affect the others), then the likelihood is given by the **binomial** distribution

$$P(K \text{ heads}, N \text{ tosses}|\ \text{bias} = H, I) = \binom{N}{K} H^K (1-H)^{N-K}.$$

THE FAIR (?) COIN – POSTERIOR(S)

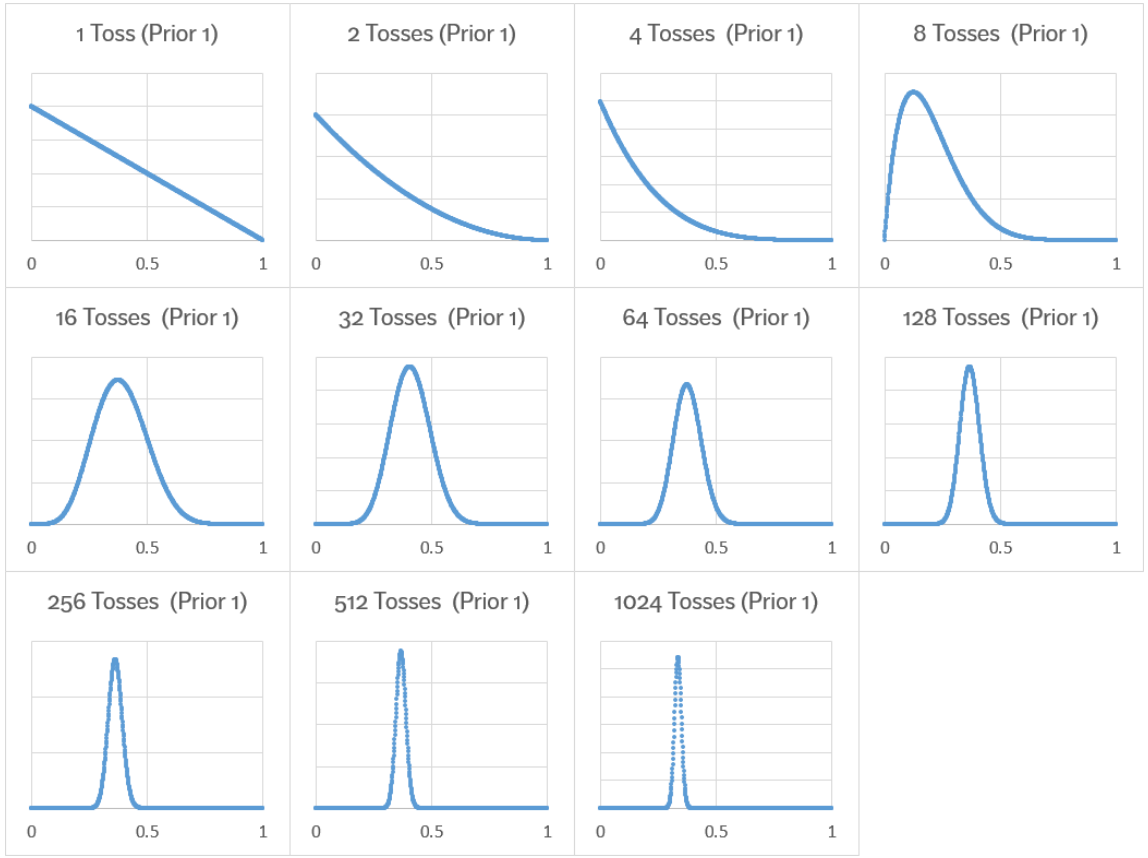
Combining both of these together, we get

$$P\left(H \,\middle|\, K \text{ heads in } N \text{ tosses}; I\right) \propto H^K (1-H)^{N-K} \times P_i(H|I),$$

where $i = 1, 2, 3, \text{or } 4$.

We should be able to estimate the bias H^* by studying the posterior distribution for each of the 4 priors, for various number of throws N.

THE FAIR (?) COIN – POSTERIORS – NON-INFORMATIVE PRIOR

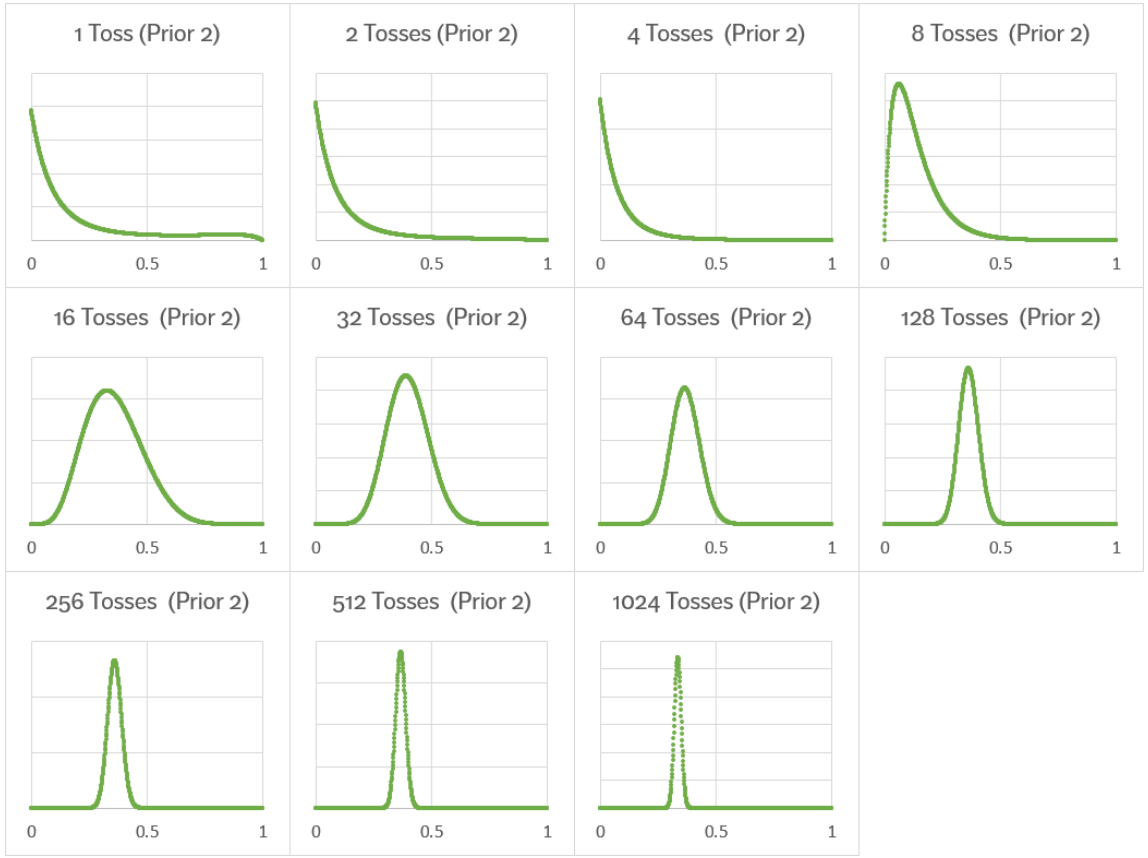


With a **non-informative** prior, the sought posterior is simply proportional to the likelihood.

Note that the central limit theorem seem to kick in after ~30 tosses.

After 128 tosses (with this specific series of tosses), we are fairly certain that the coin must be biased ($0.25 \leq H^* \leq 0.40$?)

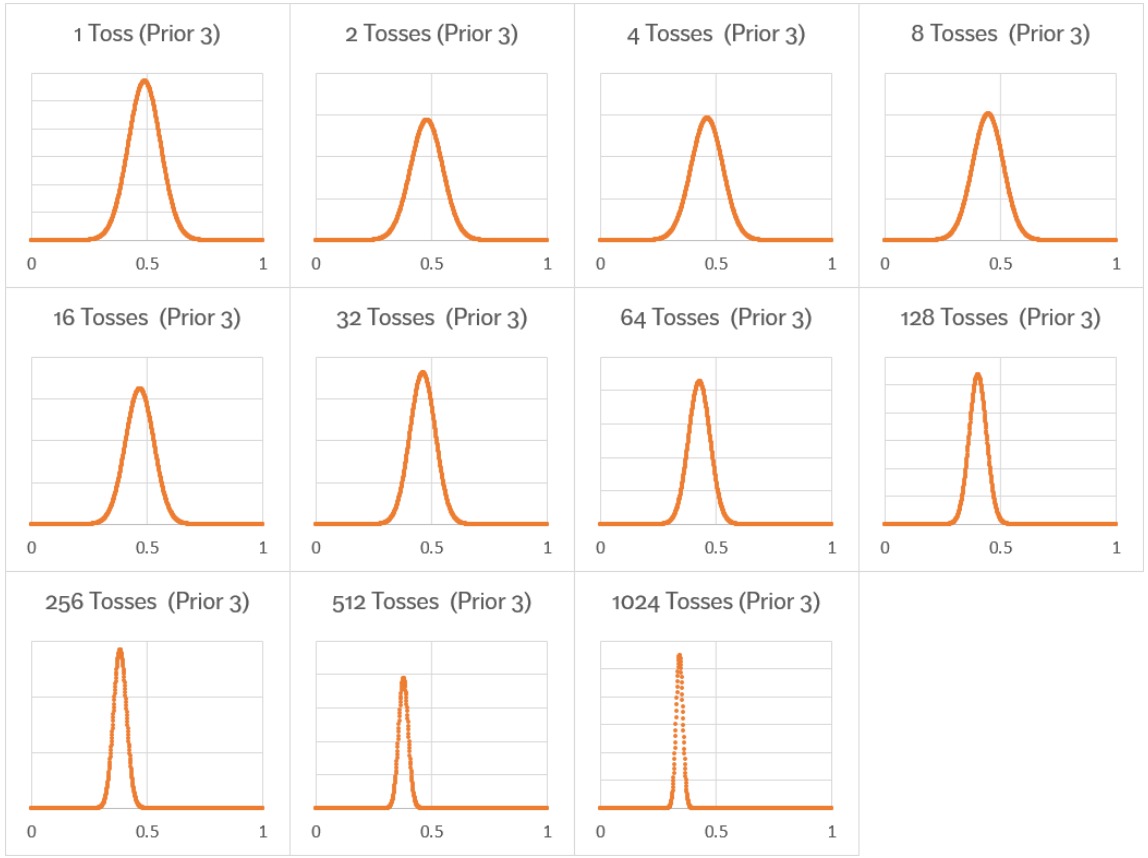
THE FAIR (?) COIN – POSTERIORS – FOUL PLAY PRIOR



With a **foul play** prior, we suspect early on that the bias is smaller than 0.5; the subsequent series of tosses moves the bias to a value $0.25 \leq H^* \leq 0.40$ fairly quickly, as was the case with the non-informative prior.

Note the shrinking of the posterior with an increasing number of tosses.

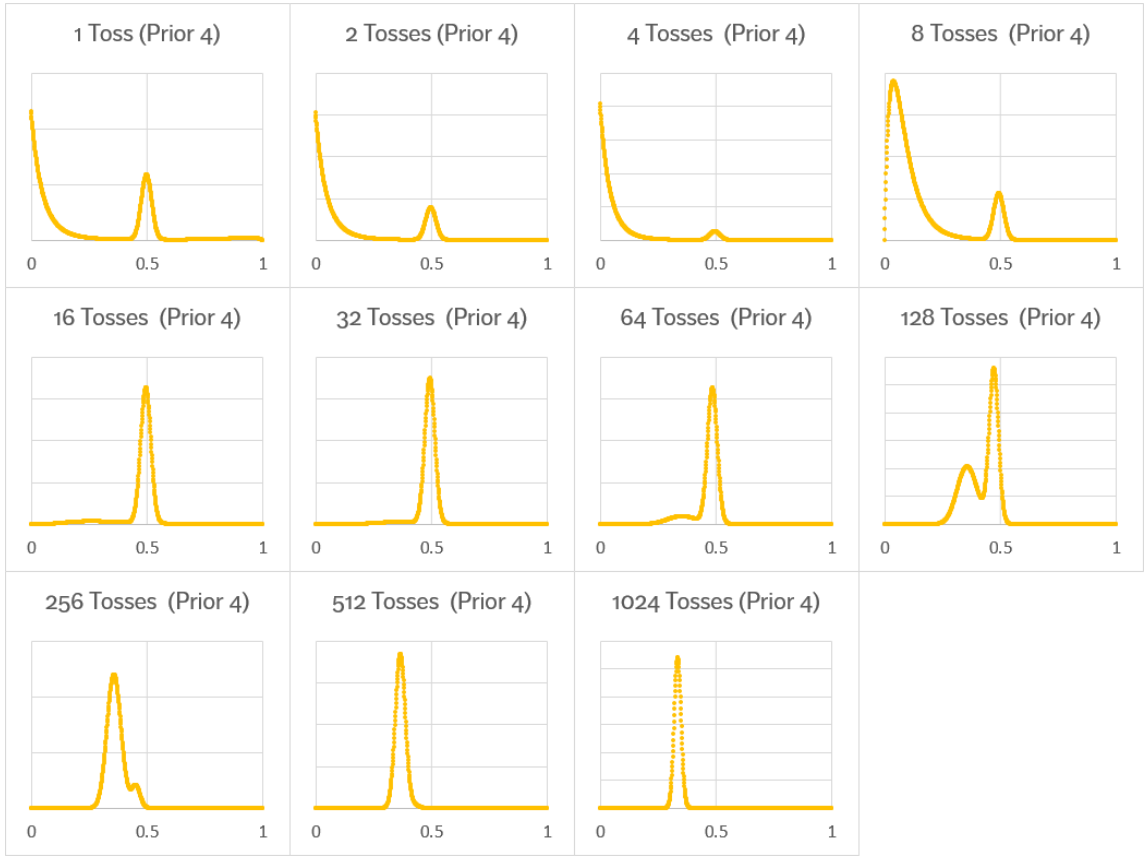
THE FAIR (?) COIN – POSTERIORS – REGULAR COIN PRIOR



With a **regular coin** prior, early results do not strongly suggest that the coin is biased (the prior gives little credence to the notion that the bias could lie in $0.25 \leq H^* \leq 0.40$).

Note the smoother convergence of the posterior.

THE FAIR (?) COIN – POSTERIORS – DOUBTFUL PRIOR



With a **doubtful** prior, the competing hypotheses compete before converging to a bias in $0.25 \leq H^* \leq 0.40$.

Note the slower convergence to a gaussian posterior.

EXAMPLE: THE SALARY QUESTION

A CURSORY GLANCE AT BAYESIAN DATA ANALYSIS

THE SALARY QUESTION – SET-UP

Income information has been collected for 4782 individuals, with demographics.

The table to the right shows some of the summary statistics for the dataset.

Question: is there a link between the demographic information and income?

Gender	Age	Edu	N	min	max
M	15-24	0	254	$ 13,970	$ 54,567
M	15-24	1	179	$ 25,871	$ 75,389
M	25-54	0	729	$ 15,560	$ 71,783
M	25-54	1	735	$ 28,329	$ 138,185
M	55-64	0	279	$ 21,966	$ 83,503
M	55-64	1	227	$ 58,384	$ 99,530
F	15-24	0	240	$ 917	$ 56,639
F	15-24	1	184	$ 20,361	$ 82,115
F	25-54	0	671	$ 14,161	$ 71,394
F	25-54	1	758	$ 22,691	$ 111,277
F	55-64	0	302	$ 20,719	$ 66,912
F	55-64	1	224	$ 53,840	$ 102,436
			4782		

THE SALARY QUESTION – SET-UP

How could you answer this question?

THE SALARY QUESTION – SET-UP

What if you had reason to suspect that reported incomes follow a (potentially) different distribution for each group?

In the Bayesian framework, you would be interested in the posterior distribution

$$P(\text{parameters}|\text{data}, i, I), i = 1, \dots, 12.$$

If we assume (for no particular good reason) that the reported incomes are **normally distributed** for each group, then we seek

$$P(\mu_i, \sigma_i|\text{reported incomes in group } i, I), i = 1, \dots, 12.$$

THE SALARY QUESTION – PRIORS

What the **priors** $P(\mu_i, \sigma_i|I), i = 1, \ldots, 12$ could be is not easy to answer. One could naively pick a joint distribution which **peaks** at the sample mean $\bar{x}_i$ and standard deviation s_i for each group i, but there are sampling design issues associated with this approach.

Why not select, instead, a prior "which expresses **complete ignorance** except for the fact that μ_i is a **location** parameter and σ_i is a **scale** parameter" [Janyes, 2003; Oliphant; 2006]. This translates into using a prior $P_1(\mu_i, \sigma_i|I) \propto {\sigma_i}^{-1}$.

For comparison's sake, we will also consider the prior $P_2(\mu_i, \sigma_i|I) \propto \mu_i^{500}{\sigma_i}^{-4}$.

THE SALARY QUESTION – PRIORS

What could those priors represent, in the real world? What happens to the probabilities when σ_i increases? When μ_i increases?

Note, as well, that these ”priors” are not normalizable over the positive quadrant in (μ, σ)-space. Instead, we only consider them over a suitable finite sub-region.

*green: high, red: low

THE SALARY QUESTION – PRIORS

$P_1(\mu_i, \sigma_i|I)$

$P_2(\mu_i, \sigma_i|I)$

THE SALARY QUESTION – LIKELIHOOD

Let's denote the number of observations in group i by N_i. The **likelihood** is the probability

$$P\left(\text{reported incomes } \{x_{k,i}\} \text{ in group } i \mid \mu_i, \sigma_i, I\right), i = 1, \ldots, 12.$$

We've assumed normality for any given observation. If we assume further that all observations are independent, then

$$P\left(\{x_{k,i}\} \mid \mu_i, \sigma_i, I\right) \propto \prod_{k=1}^{N_i} \sigma_i^{-1} \exp\left(\frac{-\left(\mu_i - x_{k,i}\right)^2}{2\sigma_i^2}\right), i = 1, \ldots, 12.$$

THE SALARY QUESTION – POSTERIOR(S)

Combining both of these together, we get

$$P_1\left(\mu_i, \sigma_i | \{x_{k,i}\}, I\right) \propto \sigma_i^{-(N_i+1)} \prod_{k=1}^{N_i} \exp\left(\frac{-\left(\mu_i - x_{k,i}\right)^2}{2\sigma_i^2}\right),$$

and

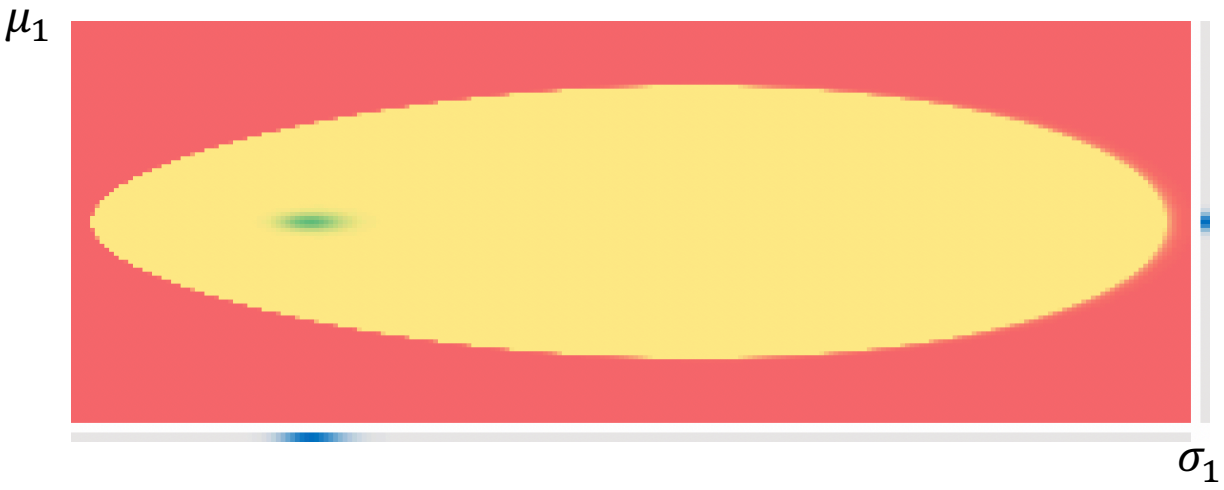
$$P_2\left(\mu_i, \sigma_i | \{x_{k,i}\}, I\right) \propto \mu_i^{500} \sigma_i^{-(N_i+4)} \prod_{k=1}^{N_i} \exp\left(\frac{-\left(\mu_i - x_{k,i}\right)^2}{2\sigma_i^2}\right),$$

for $i = 1, \ldots, 12$ over some suitable sub-region in parameter space.

*green: high, red: low

THE SALARY QUESTION – POSTERIORS – GROUP 1

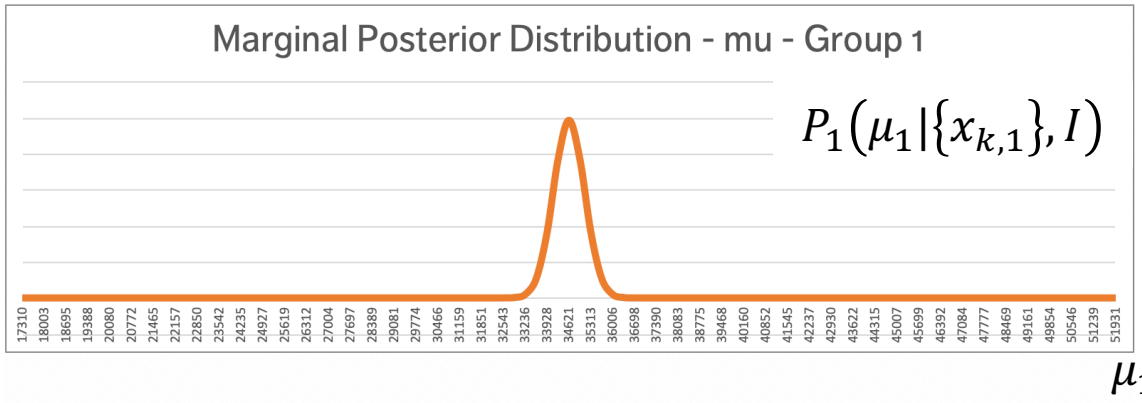
$P_1(\mu_1, \sigma_1 | \{x_{k,1}\}, I)$

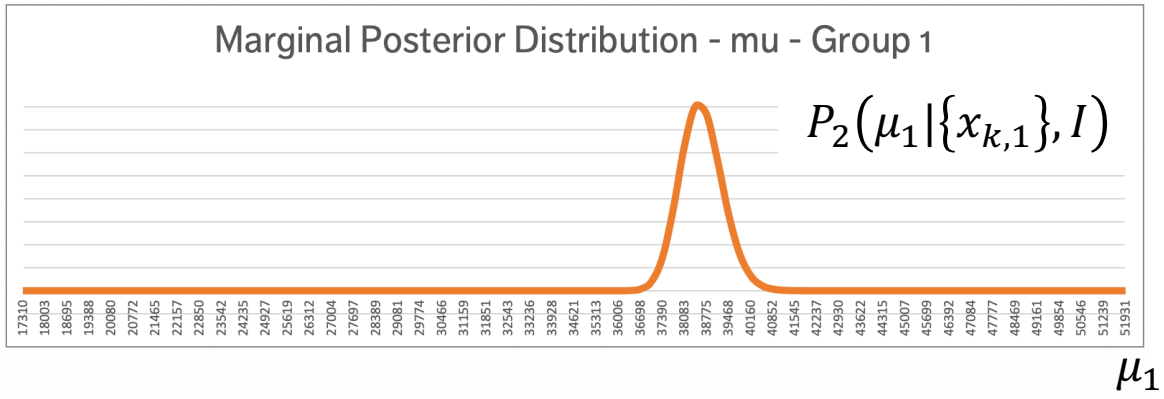


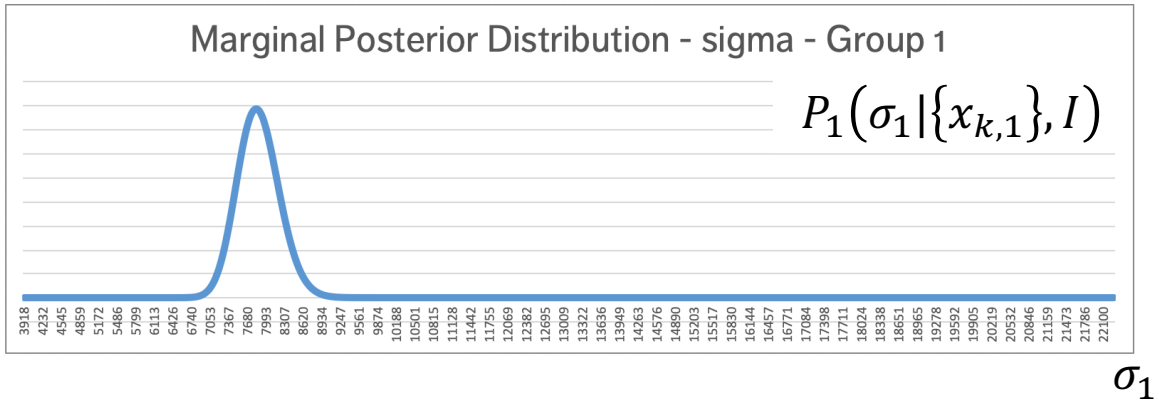
$P_2(\mu_1, \sigma_1 | \{x_{k,1}\}, I)$

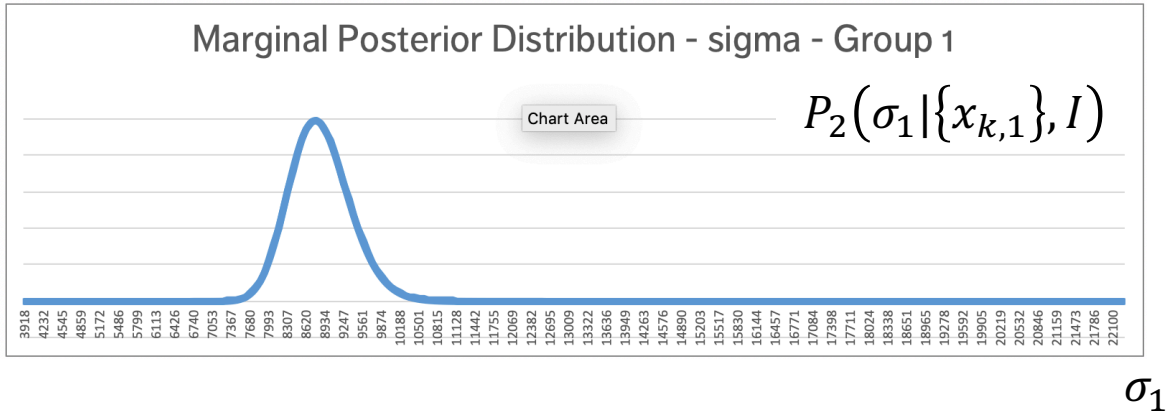
*green: high, red: low

THE SALARY QUESTION – POSTERIORS – GROUP 1









*green: high, red: low

THE SALARY QUESTION – EXERCISE

Using the Excel spreadsheet, estimate the parameters $(\mu_i, \sigma_i), i = 1, \ldots, 12$.

EXAMPLE: MONEY ($ BILL Y'ALL)

A CURSORY GLANCE AT BAYESIAN DATA ANALYSIS

MONEY ($ BILL Y'ALL) – THE SET-UP

The **question**: how many 5$ dollar bills are there in circulation?

The **problem**: we cannot count them all.

MONEY ($ BILL Y'ALL) – THE SET-UP

The **solution**: "catch and release"

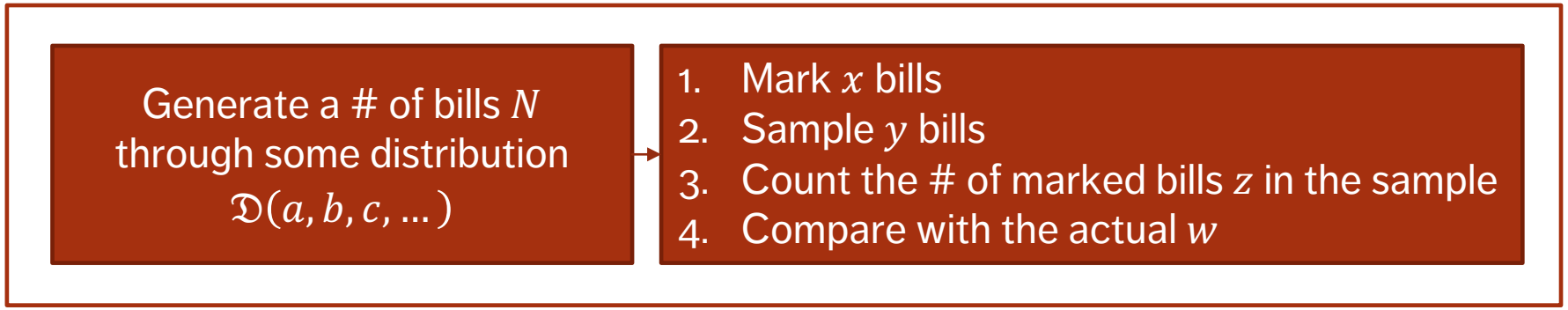
1. Capture a few 5$ bills.
2. Mark them and put them back in circulation.
3. At some later point, capture a few 5$ bills.
4. Count how many are marked.

For instance, $x = 500$ bills might have been marked initially; $y = 300$ bills might have been re-captured at stage 3, and $w = 127$ of which were marked.

What is the most probably number of bills N in circulation?

MONEY ($ BILL Y'ALL) – FITTING THE MODEL

Unlike in the previous examples when we were trying to estimate the parameters from the data using a **generative model**, in this example we are trying to estimate data from parameters.



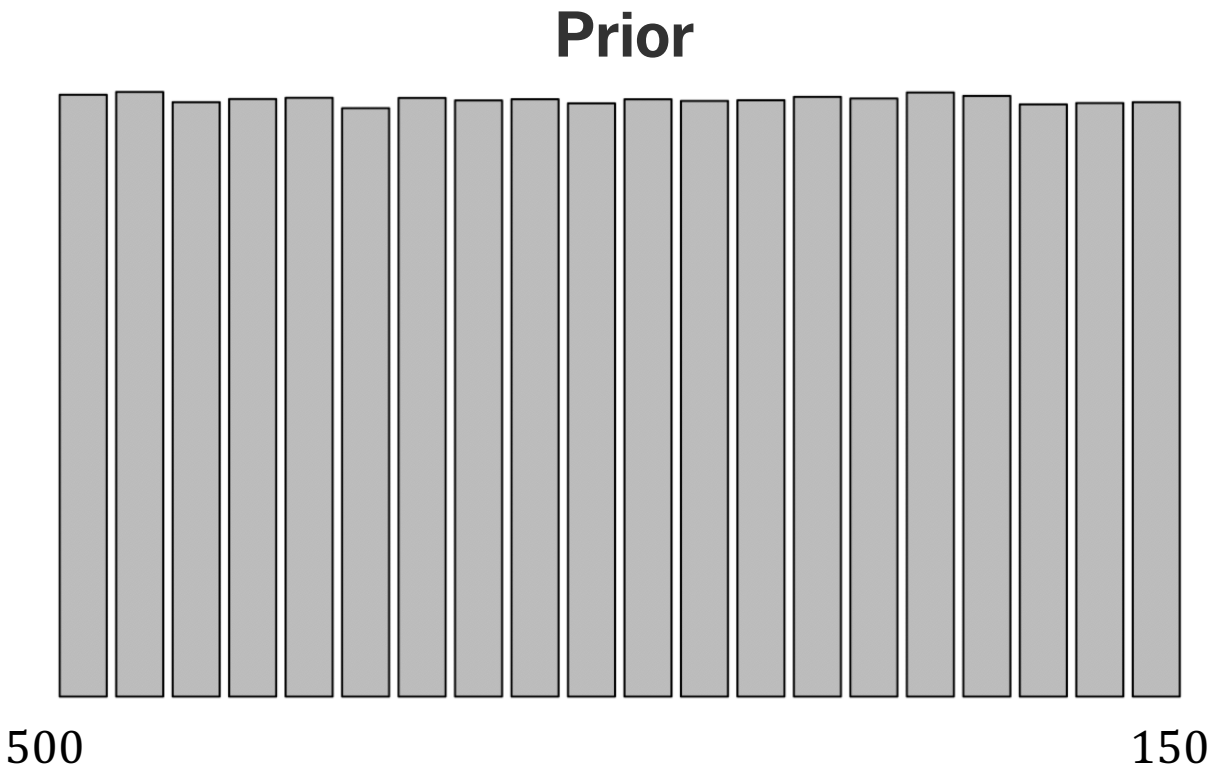
Repeat to get a distribution of z's

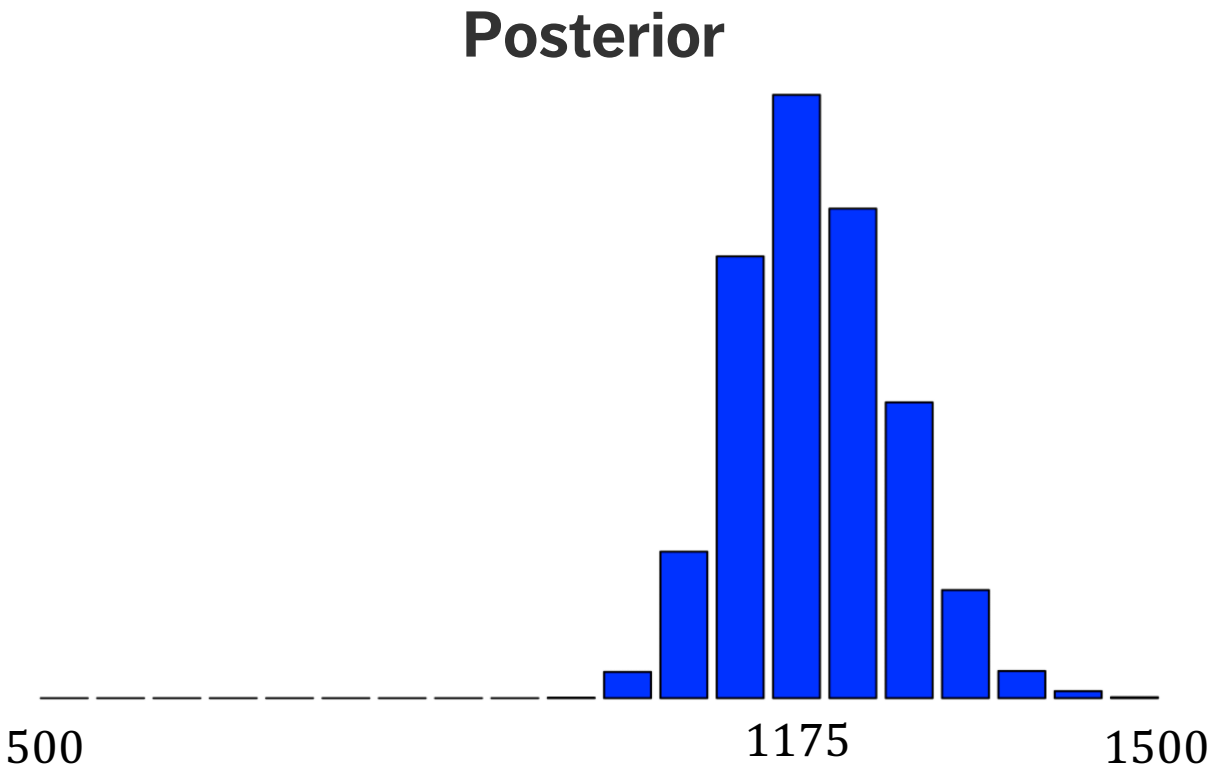
x, y, w are given; z, N to be found

MONEY ($ BILL Y'ALL) – FITTING THE MODEL (SIMPLE)

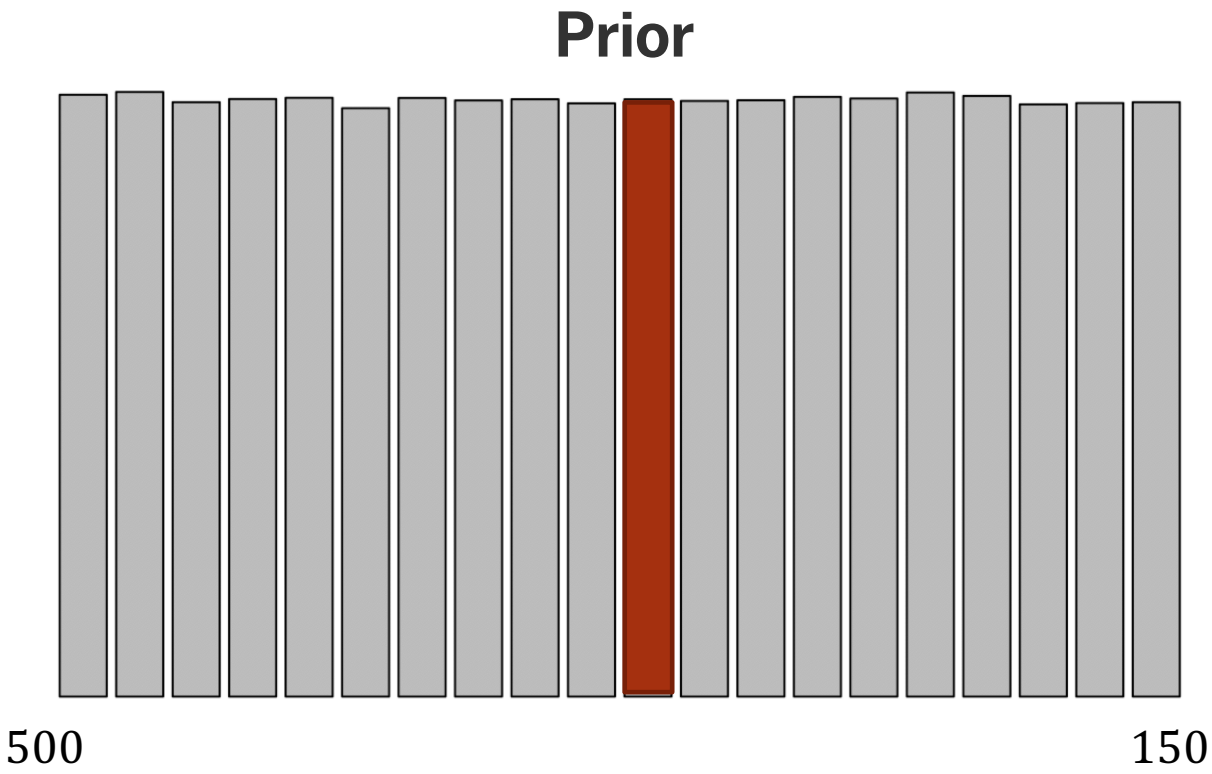
1. Draw a **large** random sample of # of bills N from an acceptable "prior" distribution on the parameters.
2. Using the N's and the generative model (with x and y given), produce a (synthetic) # of marked bills z in each sample.
3. Retain only those values of N values for which $z = w$.

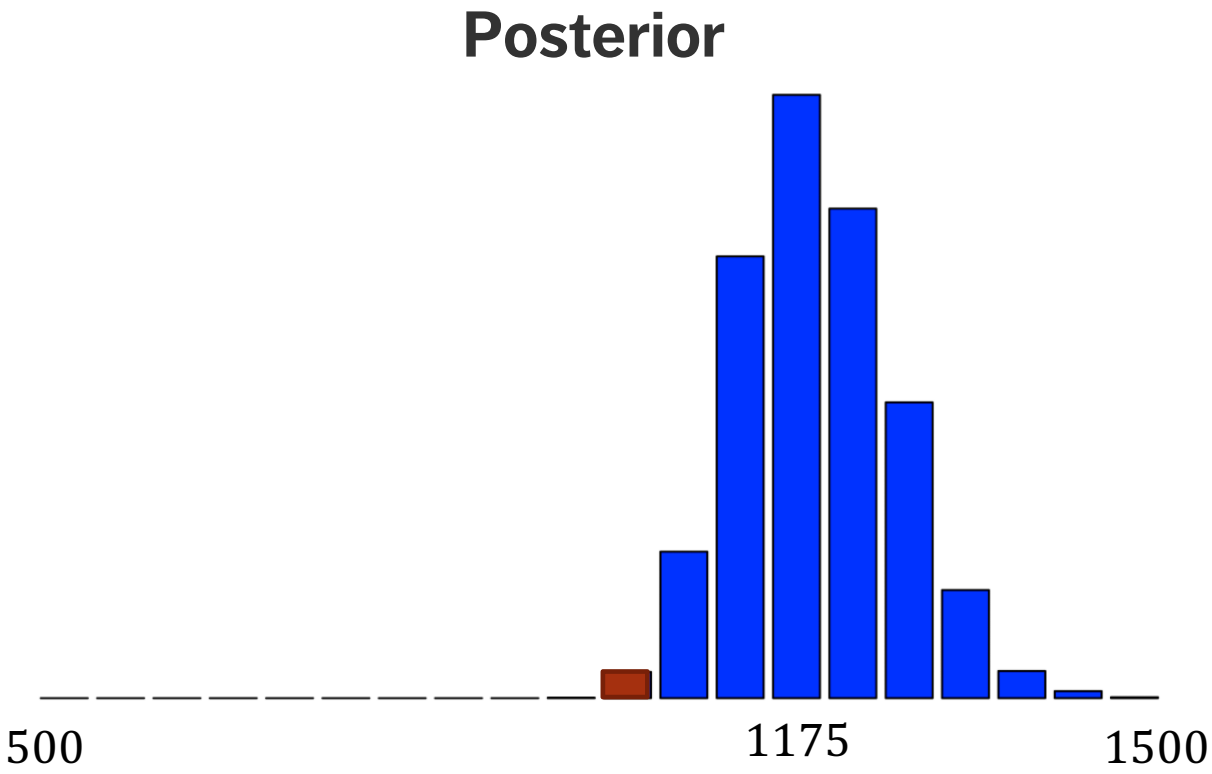
MONEY ($ BILL Y'ALL) – FITTING THE MODEL (SIMPLE)





MONEY ($ BILL Y'ALL) – FITTING THE MODEL (SIMPLE)

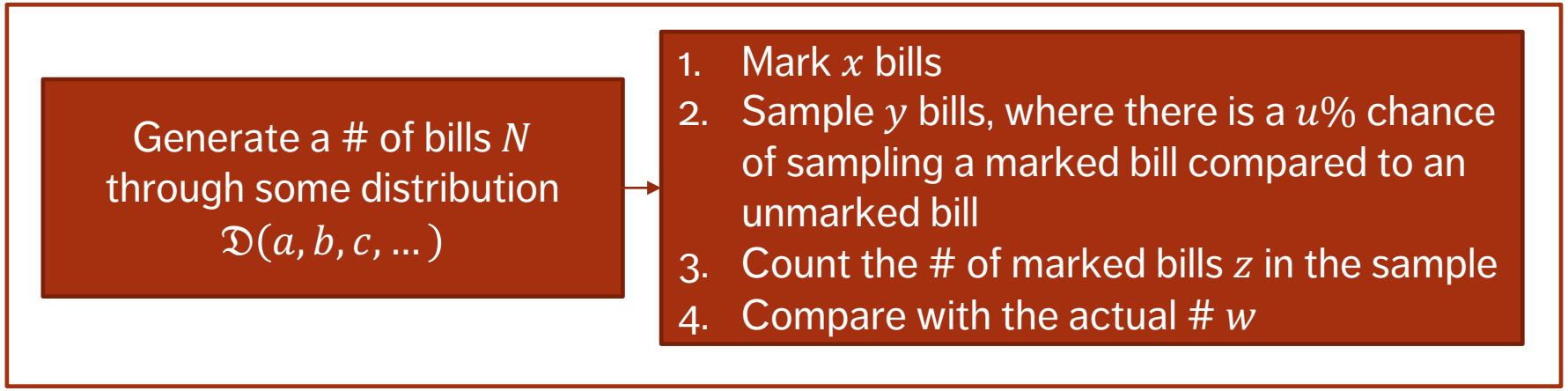




$$P(N = 1000|z = 127, I) \propto P(z = 127|N = 1000, I) \times P(N = 1000|I)$$

MONEY ($ BILL Y'ALL) – MARKED BILLS ARE BRITTLE (?)

It may be the case that the process of marking the bills might damage them somehow, so that they may be retired sooner than one would expect (with prob. 90%, say).



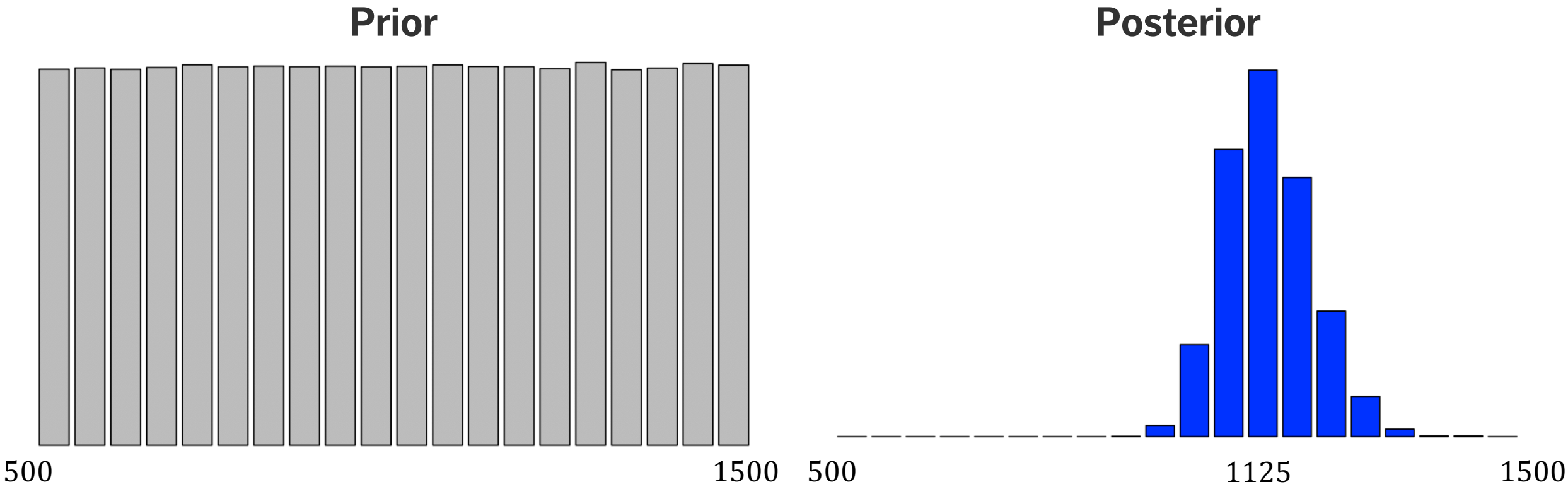
Repeat to get a distribution of z's

x, y, u, w are given; z, N to be found

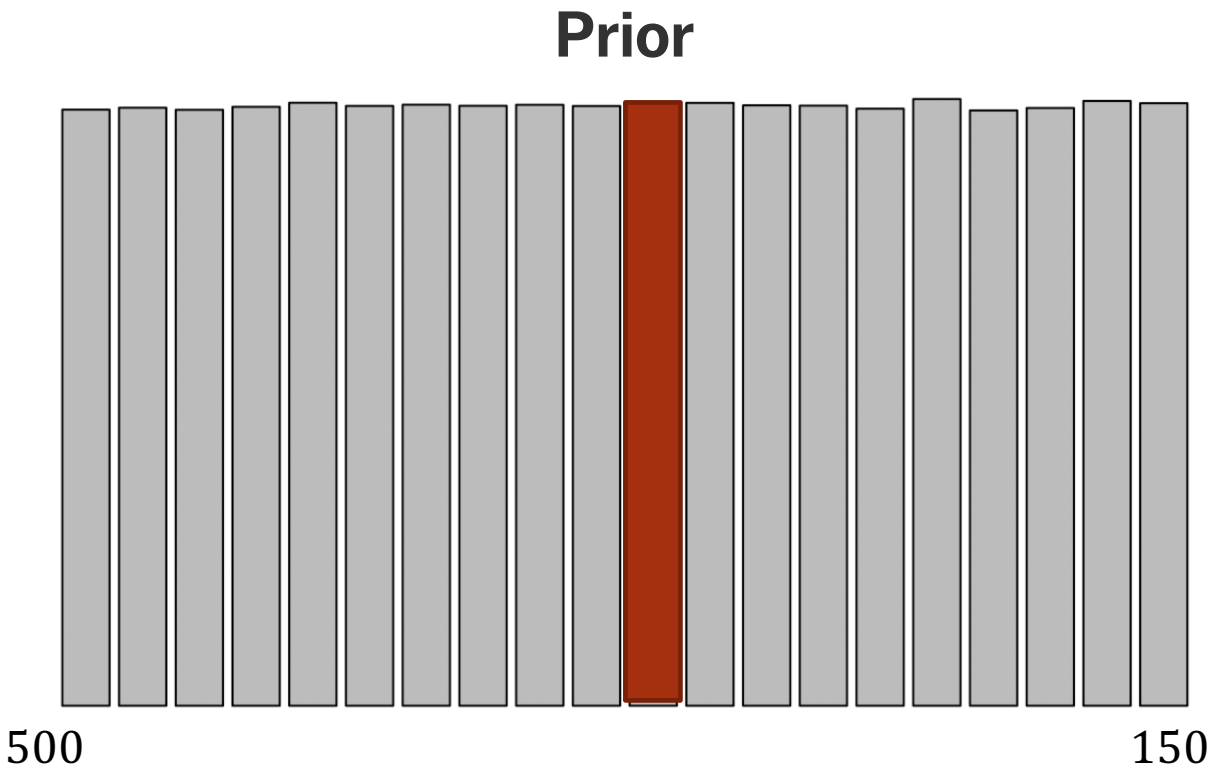
MONEY ($ BILL Y'ALL) – MARKED BILLS ARE BRITTLE (?)

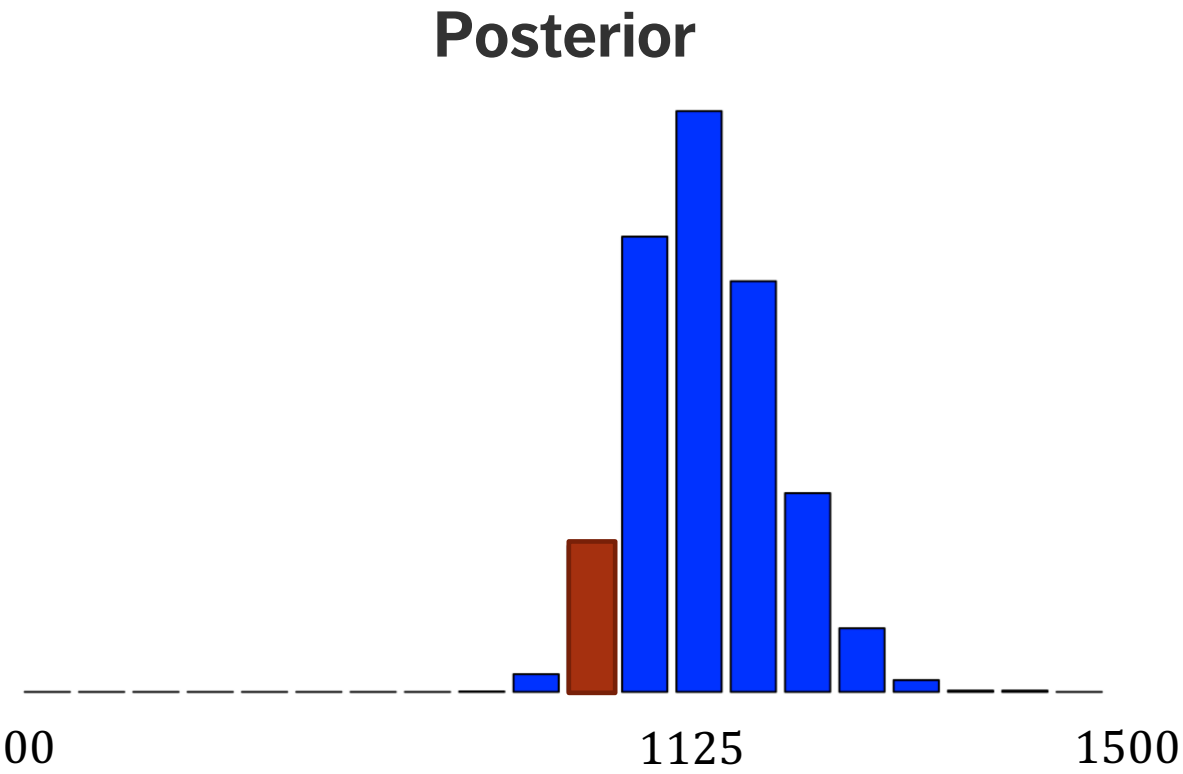
1. Draw a **large** random sample of # of bills N from an acceptable "prior" distribution on the parameters.
2. Using the N's and the generative model (with x, y and u given), produce a (synthetic) # of marked bills z in each sample.
3. Retain only those values of N values for which $z = w$.

MONEY ($ BILL Y'ALL) – MARKED BILLS ARE BRITTLE (?)



MONEY ($ BILL Y'ALL) – MARKED BILLS ARE BRITTLE (?)





$$P(N = 1000|z = 127, I) \propto P(z = 127|N = 1000, I) \times P(N = 1000|I)$$

MONEY ($ BILL Y'ALL) – LISTEN TO THE BANKER

An old banker thinks that there should be about 1000 bills in circulation. How can we incorporate this piece of information?

Generate a # of bills N through **another** distribution $\mathfrak{D}(a, b, c, \dots)$

1. Mark x bills
2. Sample y bills, where there is a $u\%$ chance of sampling a marked bill compared to an unmarked bill
3. Count the # of marked bills z in the sample
4. Compare with the actual # w

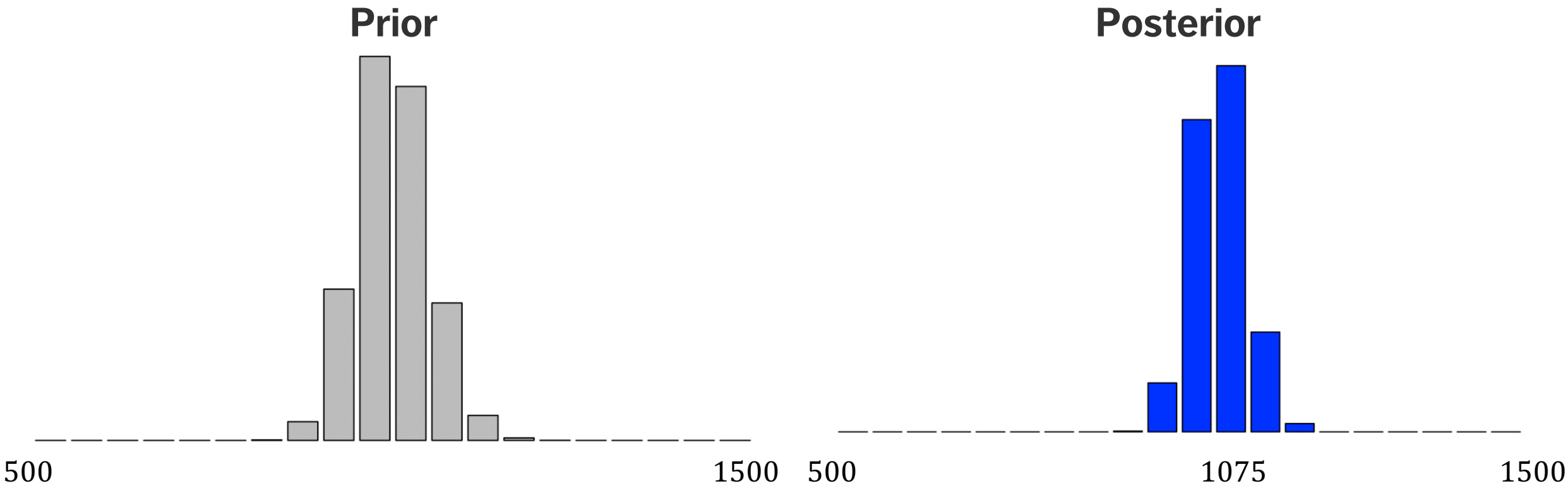
Repeat to get a distribution of z's

x, y, u, w are given; z, N to be found

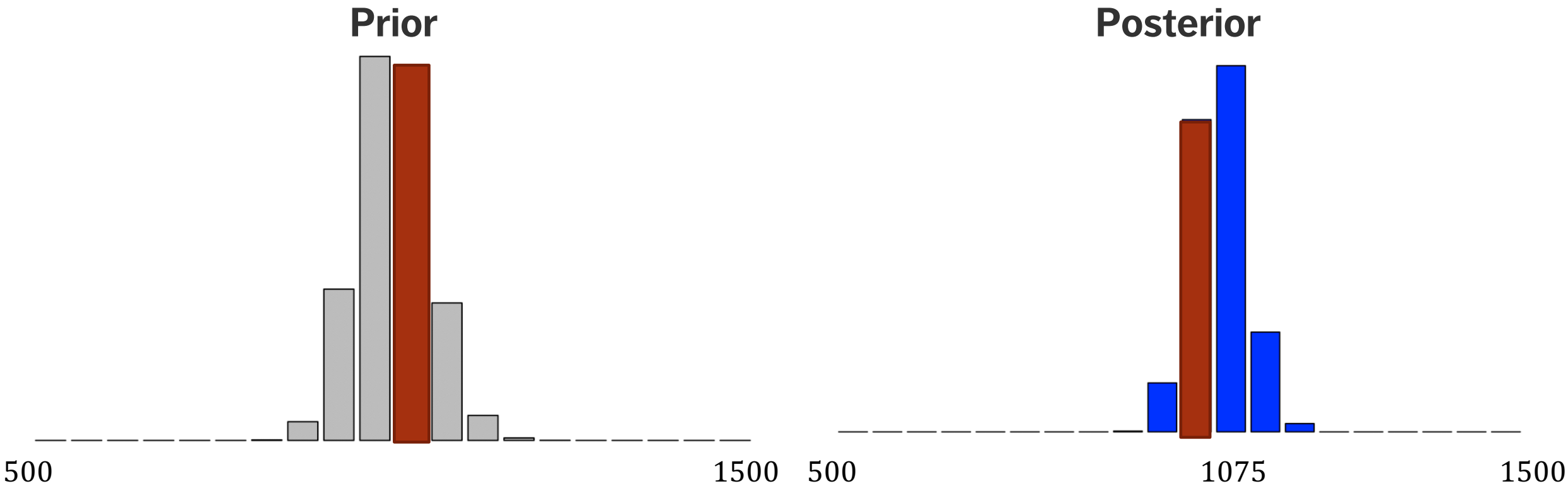
MONEY ($ BILL Y'ALL) – LISTEN TO THE BANKER

1. Draw a **large** random sample of # of bills N from a negative binomial distribution.
2. Using the N's and the generative model (with x, y and u given), produce a (synthetic) # of marked bills z in each sample.
3. Retain only those values of N values for which $z = w$.

MONEY ($ BILL Y'ALL) – LISTEN TO THE BANKER



MONEY ($ BILL Y'ALL) – LISTEN TO THE BANKER



$$P(N = 1000|z = 127, I) \propto P(z = 127|N = 1000, I) \times P(N = 1000|I)$$

OUTLINE

Part 1

1. Plausible Reasoning
2. The Rules of Probability
3. Bayes' Theorem
4. Example: the Fair (?) Coin
5. Example: the Salary Question
6. Example: Money ($ Bill Y'All)

Part II

7. Marginalization (coming soon)
8. Prior Distributions
9. Model Selection (coming soon)
10. Naïve Bayes (coming soon)
11. Bayesian Inference (coming soon)
12. MCMC and Numerical Methods

PRIOR DISTRIBUTIONS

A CURSORY GLANCE AT BAYESIAN ANALYSIS

PRIOR DISTRIBUTIONS

Specifying a model necessarily means **providing a prior** distribution for the unknown parameters.

Prior plays a crucial role in Bayesian inference through the **updating statement**

$$p(\theta|D) \propto p(\theta) \times p(D|\theta)$$

Choice of prior is **subjective** (decision to use a prior is left entirely up to the analyst).

PRIOR DISTRIBUTIONS

But the choice of prior is **no more subjective than the choice of likelihood, the selection of a sample, the estimation framework, or the statistic used for data reduction**.

Choice of prior can affect posterior conclusions, in particular when the sample size is small.

CONJUGATE PRIORS

In general: posterior distribution for vector θ has no simple analytical representation.

Posterior distributions must be **estimated numerically** (not exact).

Exceptions: **conjugate priors**

- joint property of a prior and a likelihood $\Longrightarrow$ posterior has same form as prior (but with different parameters)

CONJUGATE PRIORS

Likelihood	Prior	Hyperparameters
Bernouilli	Beta	$\alpha > 0, \beta > 0$
Binomial	Beta	$\alpha > 0, \beta > 0$
Poisson	Gamma	$\alpha > 0, \beta > 0$
Normal for μ	Normal	$\mu \in \mathbb{R}, \sigma^2 > 0$
Normal for σ^2	Inverse Gamma	$\alpha > 0, \beta > 0$
Exponential	Gamma	$\alpha > 0, \beta > 0$

Example:

- likelihood = Binomial $P(s, n|q) = \binom{n}{s} q^s (1-q)^{n-s}$, prior = Beta$(\alpha, \beta)$
- posterior = Beta$(\alpha + s, \beta + n - s)$

CONJUGATE PRIORS

Conjugate priors are mathematically convenient, and they can be quite flexible, depending on the specific hyperparameters we use; **but they reflect very specific prior knowledge and should be eschewed unless we truly possess that prior knowledge**.

Alternatives:

- uninformative priors
- informative priors
- maximum entropy (MaxEnt) priors

UNINFORMATIVE PRIORS

Uninformative priors intentionally provide very little specific information about the unknown parameter(s).

Rationale: 'to let the data speak for itself,' so that inferences are unaffected by information external to the current data.

Classic Example: uniform prior

- for data following a Bernoulli(θ) distribution, a uniform prior on θ is $P(\theta) = 1$ on $0 \leq \theta \leq 1$.
- for data following a normal $N(\mu, 1)$ distribution, the uniform prior on the support of μ is **improper** however, such a choice could still be acceptable as long as the resulting posterior is normalizable.

INFORMATIVE PRIORS

Informative priors are those that **deliberately** insert information that researchers have at hand into the analysis.

Reasonable since prior scientific knowledge should play a role in statistical inference.

2 important requirements:

- overt declaration of prior specification
- detailed sensitivity analysis to show the effect of these priors relative to uninformed types.

Transparency is required to avoid the common pitfall of **data fishing**; sensitivity analysis can provide a sense of exactly how informative the prior is.

INFORMATIVE PRIORS

Where do informative priors come from, in the first place?

- past studies, published work, researcher intuition
- interviewing domain experts
- convenience with conjugacy
- non-parametric and other data-derived sources.

Prior information from past studies need not be in agreement.

Useful strategy: construct prior specifications from **competing school-of-thoughts** to contrast resulting posteriors and produce informed statements about the relative strength of each of them.

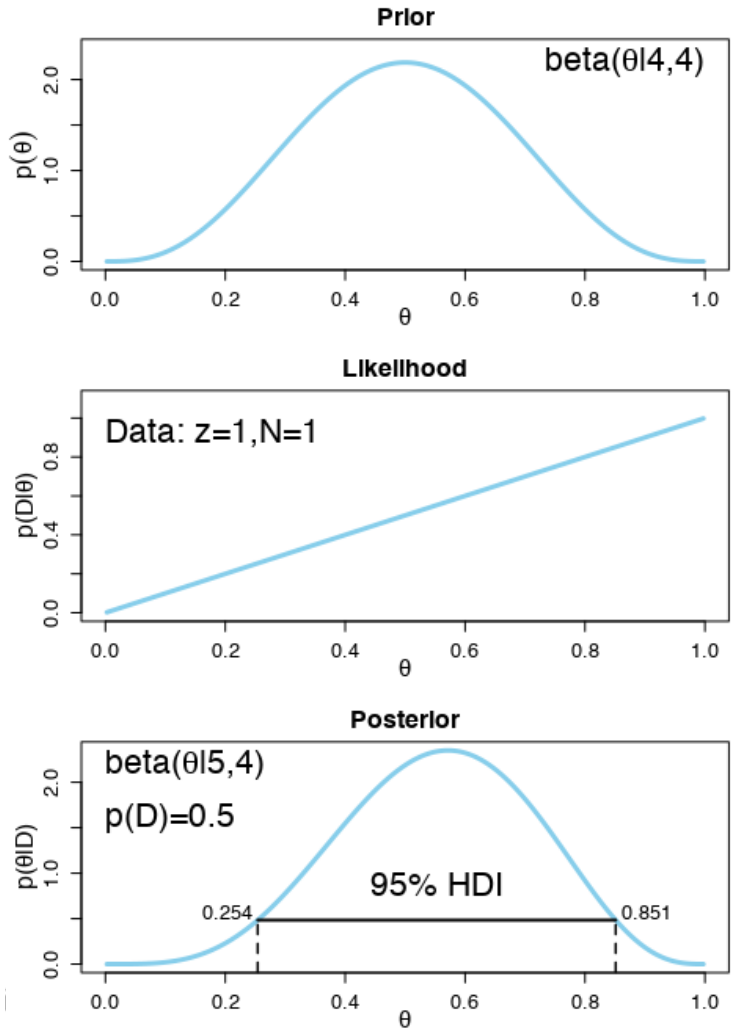
INFORMATIVE PRIORS

Example: Bernouilli likelihoods and Beta priors form conjugate priors.

1. Start with a prior distribution that expresses some uncertainty that a coin is fair: $\text{Beta}(\theta\,|4,4)$ – a coin was recorded with 4H in 8 tosses, perhaps. Flip the coin once; assume that H is obtained. What is the posterior distribution of the uncertainty in the coin's fairness θ?

   Solution: use `post=BernBeta(c(4,4),c(1))` from `Example 4.R`

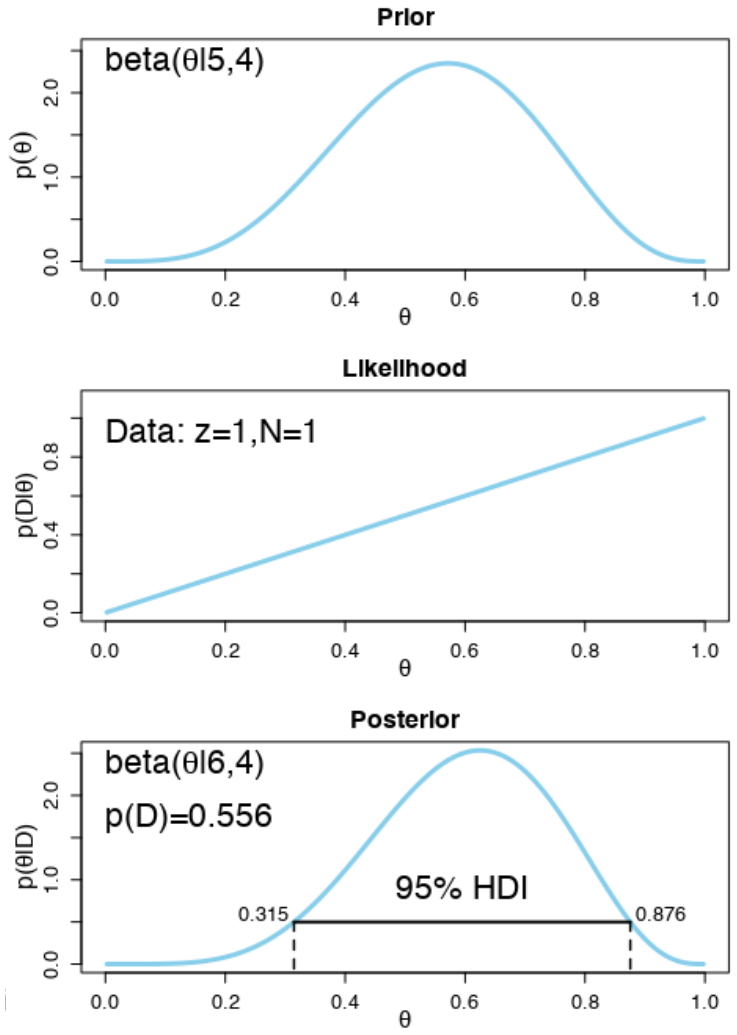
INFORMATIVE PRIORS



2. Use the posterior parameters from the previous flip as the prior for the next flip. Suppose we flip again and get a H. What is the new posterior on the uncertainty in the coin's fairness θ?

Solution: use `post=BernBeta(post,c(1))` from `Example 4.R`

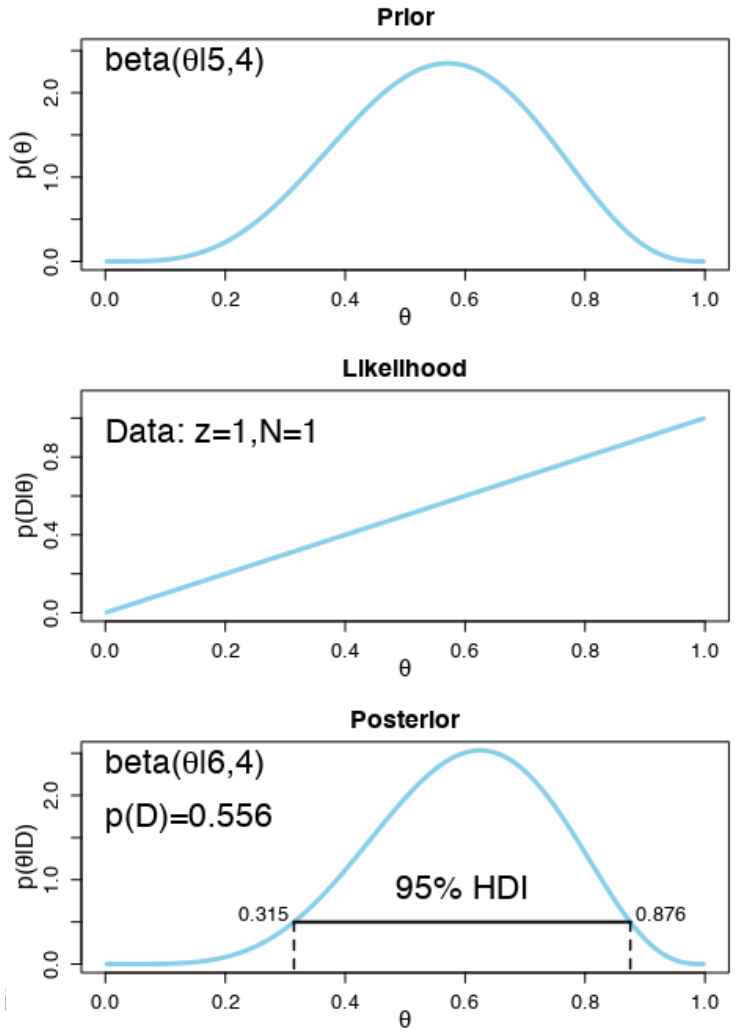
INFORMATIVE PRIORS



3. Using the most recent posterior as the prior for the next flip, flip a third time and obtain yet again a H. What is the new posterior?

Solution: in this case, we know that the posterior for the coin's fairness θ follows a $\text{Beta}(\theta\,|7,4)$ distribution. Does 3H in a row give you pause? Is there enough evidence to suggest that the coin is not fair? What if you flipped 18 H in a row from this point on?

INFORMATIVE PRIORS



4. Suppose that a friend has a coin that we know comes from a magic store; as a result, we believe that the coin is strongly biased in either of the two directions (it could be a trick coin with both sides being H, for instance), but we don't know which one it favours. We will express the belief of this prior as a Beta distribution.

   Let's say that our friend flips the coin five times; resulting in 4H and 1T. What is the posterior distribution of the coin's fairness θ?

MAXIMUM ENTROPY PRIORS

Whether the priors are uninformative or informative, we search for the distribution that best encodes the prior state of knowledge from a set of **trial** distributions.

Let X be discrete, of cardinality M, with probability density $p(X) = (p_1, \ldots, p_M)$. The **entropy** $H(p)$ of p is

$$H(p) = -\sum_{i=1}^{M} p_i \log p_i \,, \text{ with } 0 \times \log 0 = 0$$

MAXIMUM ENTROPY PRIORS

The **maximum entropy principle** (MaxEnt) states:

> given a class of trial distributions with constraints, the optimal prior is the trial distribution with the **largest entropy**.

As an example, the most basic constraint is for p to lie in the probability simplex, that is, $p_1 + \cdots + p_M = 1$ and $p_i \geq 0$ for all i.

With those basic constraints, the maximum entropy prior is the **uniform distribution**.

MCMC AND NUMERICAL METHODS

A CURSORY GLANCE AT BAYESIAN ANALYSIS

POSTERIOR DISTRIBUTIONS

Posteriors are used to estimate a variety of model parameters of interest:

- mean, median, mode, etc.

It is possible to construct **credible intervals/regions** directly from the posterior.

Because the posterior is a full distribution on the parameters, it is possible to make all sorts of probabilistic statements about their values, such as:

- "I am 95% sure that the true parameter value is bigger than 0.5"
- "There is a 50% chance that θ_1 is larger than θ_2"
- etc.

POSTERIOR DISTRIBUTIONS – HDI

The best approach is to build a $1 - \alpha$ credible interval of θ-values using the **highest density interval** (HDI):

- a region C_k in the parameter space
- $C_k = \{\theta : p(\theta|D) \geq k\}$
- k is the largest number such that $\int_{C_k} p(\theta|D)\, d\theta = 1 - \alpha$

The value k is the height of a horizontal line (or hyperplane, in the case of multivariate posteriors) overlaid on the posterior and for which the area under the curve bounded by the intersections with the hyperplane is $1 - \alpha$. In most cases, k can be found **numerically**.

POSTERIOR DISTRIBUTIONS – HDI

Example: It is an election year and you are interested in knowing whether the general population prefers candidate A or candidate B. A recently published poll states that of 400 randomly sampled people, 232 preferred candidate A, while the remainder preferred candidate B.

1. Suppose that before the poll was published, your prior belief was that the overall preference follows a uniform distribution. What is the 95% HDI on your belief after learning of the poll result?

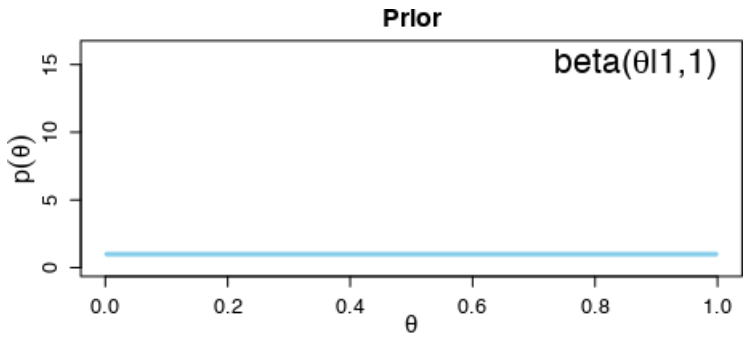
   (**Hint:** what parameters would you use in `BernBeta()`?)

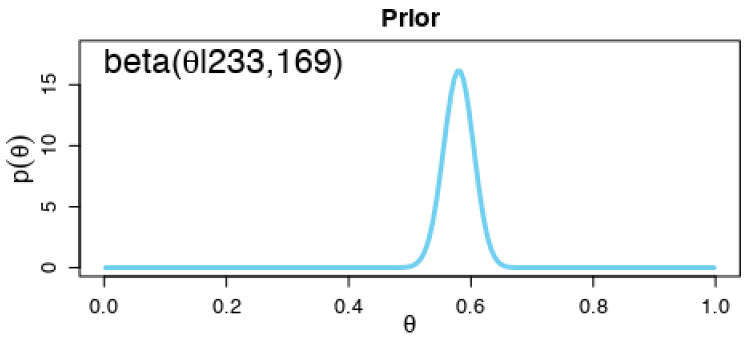
POSTERIOR DISTRIBUTIONS – HDI

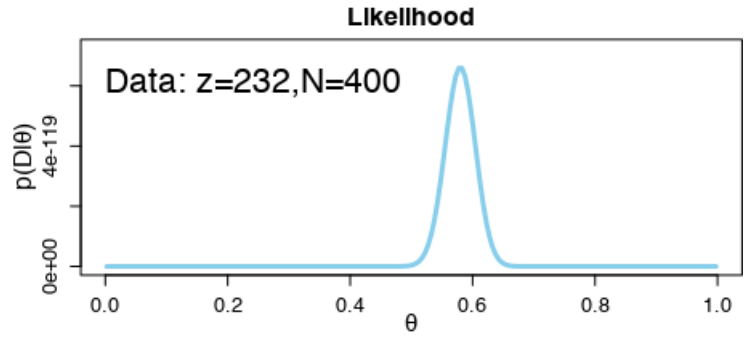
2. Based on the poll, is it credible to believe that the population is equally divided in its preferences among candidates?

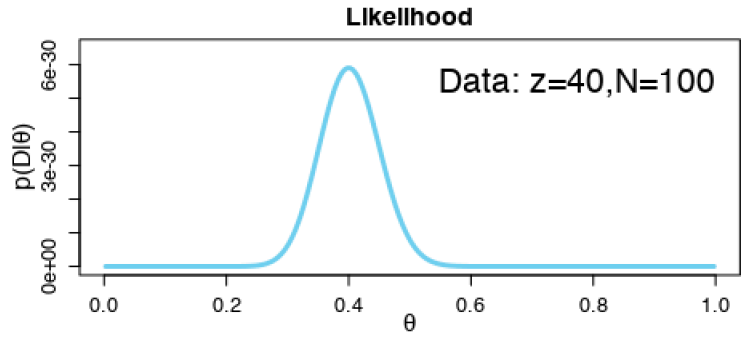
3. Assume that a subsequent poll of 100 individuals is published. How many people would have to say that they prefer candidate B for you to change the answer you gave in 2.?

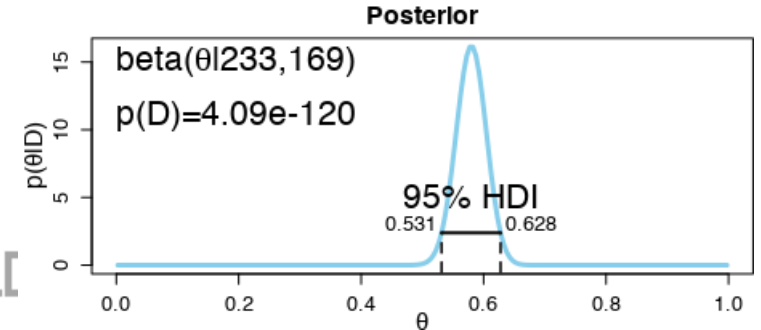
POSTERIOR DISTRIBUTIONS – HDI

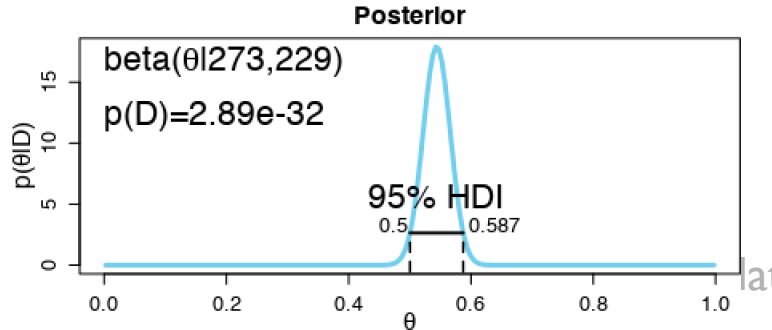












MCMC METHODS

When posteriors that cannot be manipulated analytically, it is usually possible to recreate a synthetic (or **simulated**) set of values that share the properties of the posterior (**Monte Carlo simulations**).

A **Markov chain** is an ordered, indexed set of random variables (a stochastic process) in which the values of the quantities at a given state depends probabilistically only on the values of the quantities at the preceding state.

Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on the construction of a Markov chain with the desired distribution as its equilibrium distribution.

MCMC METHODS

MCMC techniques are often applied to solve **integration** and optimization problems in large-dimensional spaces. For instance, given variables $\theta \in \Theta$ and data D, the following (typically intractable) integration problems are central to Bayesian inference:

- **normalisation:** $p(\theta|D) = \frac{p(\theta)p(D|\theta)}{\int p(\theta)p(D|\theta)d\theta}$
- **marginalisation:** $p(\theta|D) = \int p(\theta, x|D)dx$
- **expectation:** $E[f(\theta)] = \int_{\Theta} f(\theta)p(\theta|D)\, d\theta$ for functions of interest (i.e. $f(\theta) = \theta$ (mean), or $f(\theta) = (\theta - E[\theta])^2$ (variance)).

METROPOLIS-HASTINGS (MH) ALGORITHM

MH is a specific type of MCMC; it generates a **random walk** (a succession of posterior samples) so that each step in the walk is **completely independent** of the preceding steps; the decision to reject or accept the proposed step is also independent of the walk's history.

MH uses a candidate or **proposal distribution** for the posterior and constructs a Markov Chain by proposing values from this candidate distribution, and then either accepting or rejecting this value (with a certain probability).

The proposal distributions can be nearly anything, but in practice it is recommended that (really) simple ones be selected: a **normal** if the parameter of interest can be any real number (e.g. μ) or a **log-normal** if it has positive support (e.g., σ^2).

METROPOLIS-HASTINGS (MH) ALGORITHM

Algorithm 1: Metropolis-Hastings Algorithm

Initialize $x^{(0)} \sim q(x)$

for $i = 1, 2, \cdots$ **do**

  Propose: $x^* \sim q(x^{(i)}|x^{(i-1)})$

  Acceptance Probability:

$$\alpha(x^*|x^{(i-1)}) = min\left\{1, \frac{q(x^{(i-1)}|x^*)\pi(x^*)}{q(x^*|x^{(i-1)})\pi(x^{(i-1)})}\right\}$$

  $u \sim U(0, 1)$

  if $u < \alpha$ **then**

    Accept the proposal: $x^{(i)} \leftarrow x^*$

  else

    Reject the proposal: $x^{(i)} \leftarrow x^{(i-1)}$

  end

end

METROPOLIS-HASTINGS (MH) ALGORITHM

Example: go through Example 9, pp. 10-12 in *A Soft Introduction to Bayesian Data Analysis (DRAFT)*. The R file is provided in `Example 9.R`

BAYESIAN A/B TESTING

Example: go through Example 12, pp. 14-15 in *A Soft Introduction to Bayesian Data Analysis (DRAFT)*. The R file is provided in `Example 12.R`

REFERENCES

A CURSORY GLANCE AT BAYESIAN DATA ANALYSIS

REFERENCES

Sivia, D.S., Skilling, J. [2006], *Data Analysis: A Bayesian Tutorial* (2nd ed.), Oxford Science.

Silver, N. [2012], *The Signal and the Noise*, Penguin.

Jaynes, E.T. [2003], *Probability Theory: the Logic of Science*, Cambridge Press.

Kruschke, J.K. [2011], *Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan* (2nd ed.), Academic Press

Barber, D. [2012], *Bayesian Reasoning and Machine Learning*, Cambridge Press.

Gelman, A., Carloin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. [2013], *Bayesian Data Analysis* (3rd ed.), CRC Press.

Bååth, R. [2015], *Introduction to Bayesian Data Analysis with R*, UseR!

Oliphant, T.E. [2006], A Bayesian perspective on estimating mean variance, and standard-deviation from data, All Faculty Publications 278, BYU.