Bayesian Data Analysis 26

by **Patrick Boily** and **Ehssan Ghashim**

Bayesian analysis is sometimes maligned by data analysts, due in part to the perceived element of arbitrariness associated with the selection of a meaningful prior distribution for a specific problem and the (formerly formidable) difficulties involved with producing posterior distributions for all but the simplest situations.

On the other hand, it has been said that "while classical data analysts need a large bag of clever tricks to unleash on their data, Bayesians only ever really need one." With the advent of efficient numerical samplers, modern data analysts cannot shy away from adding the Bayesian arrow to their quiver.

In this chapter, we introduce the basic concepts underpinning Bayesian analysis, and we present a small number of examples that illustrate the strengths of the approach.

26.1 Plausible Reasoning

> "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

Consider the following scenario [37]: while walking down a deserted street at night, you hear a security alarm, look across the street, and see a store with a broken window, from which a person wearing a mask crawls out with a bag full of smart phones.

The natural reaction might be to conclude that the person crawling out of the store is stealing merchandise from the store.

It might be the natural reaction, but how do we actually come to this conclusion? It **cannot** come from a **logical deduction based on evidence**.[1]

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. Perhaps the owner then went into the store to retrieve items before they could be stolen, which is when you happened unto the scene.

[1] Such as would be used in mathematical 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 true

A is more plausible (why?)

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

A is false

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

B is less plausible (why?)

Figure 26.1: Deductive (left) vs. inductive (right) syllogisms.

But while the original reasoning process is not **deductive**, it is at least **plausible**, which in the logical context is called **inductive**.

We might also want to use a **weaker version** of inductive reasoning: let us say that we know that when A is true, then B is more plausible, and we also know that B is true. Then, we conclude that A is more plausible.

In the scenario described at the start of the section, 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). As you do "see them crawling out of the store with a bag of phones" (B is true), you would therefore not be surprised to find out that "the person is a thief" (A is plausible).

In **deductive reasoning**, we work from a cause to possible consequences; in **inductive reasoning**, we work from observations to possible causes.

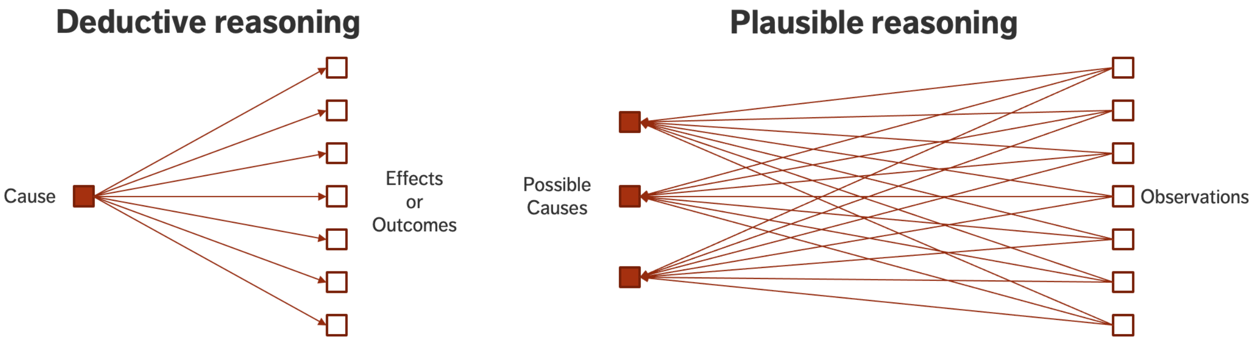


Figure 26.2: Deductive (left) vs. inductive (right) reasoning.

Plausibility relies on the notion of "surprise". In Tom Stoppard's 1966 play *Rosencrantz and Guildenstern are Dead* [320], Rosencrantz flips 92 heads in a row. This result is of course not impossible, but is it plausible? If this happened to you, what would you conclude?

26.1.1 Rules of Probability

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

In 1763, Thomas Bayes [321] published a paper on the problem of induction, that is, on arguing from the specific to the general. In modern language and notation, Bayes wanted to use binomial data comprising r

successes out of n attempts to learn about the underlying chance $\boldsymbol{\theta}$ of each attempt succeeding. Bayes' key contribution was to use a probability distribution to represent uncertainty about $\boldsymbol{\theta}$. This distribution represents **epistemiological** uncertainty, due to lack of knowledge about the world, rather than **aleatory** (random) probability arising from the essential unpredictability of future events, as may be familiar from games of chance.

In this framework, a **probability** (plausibility) represents a 'degree-of-belief' about a proposition; the probability of an event will be recorded differently by two different observers, based on the respective background information to which they have access. This **Bayesian position** was the commonplace view of probabilities in the late 1700s and early 1800s, a view shared by such luminaries as Bernoulli and Laplace.[2]

Subsequent scholars found this vague and subjective,[3] and they redefined the probability of an event as its **long-run relative frequency**, given infinite repeated trials (the so-called **frequentist position**).

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,[4] but how can we be certain that the **degree-of-belief** is a well-defined concept?

As it happens, there is a well-defined way to determine the rules of probability, based on a small list of axioms [37, 322]:

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.

In what follows, we let I denote relevant background information; X, Y, and Y_k denote various propositions, and $-X$ or $\overline{X}$ denote the negation of proposition X.

The **plausibility** of X given I is denoted by $P(X \mid I)$; it is a real number whose value can range from 0 (**false**) to 1 (**true**). The rules of probability are quite simple:

- **Sum Rule:** for all propositions X, $P(X \mid I) + P(-X \mid I) = 1$;
- **Product Rule:** for all X, Y, $P(X, Y \mid I) = P(X \mid Y; I) \times P(Y \mid I)$.

[2] Modern Bayesian statistics is still based on formulating probability distributions to express uncertainty about unknown quantities. These can be underlying parameters of a system (induction) or future observations (prediction). **Bayesian statistics** is a system for describing epistemiological uncertainty using the mathematical language of probability; **Bayesian inference** is the process of fitting a probability model to a set of data and summarizing the result with a probability distribution on the parameters of the model and on unobserved quantities (such as predictions).

[3] How can you be sure that my degree-of-belief matches yours?

[4] 92 heads in a row must mean that that the coin is biased, right?

From these two rules, we can also derive two useful corollaries:

- **Bayes' Theorem:** $P(X \mid Y; I) \times P(Y \mid I) = P(Y \mid X; I) \times P(X \mid I)$ (see next section);
- **Marginalization Rule:** $P(X \mid I) = \sum_k P(X, Y_k \mid I)$, where $\{Y_k\}$ are exhaustive and disjoint.[5]

For continuous variables, the marginalization rule becomes

$$P(X \mid I) = \int_{\Omega(Y)} P(X, Y \mid I)\, dY.$$

The **conditional probability** of A given B, $P(A \mid B)$ is the probability of A taking place given that another event B has occurred:

$$P(A \mid B; I) = \frac{P(A, B \mid I)}{P(B \mid I)} = \frac{P(A \cap B \mid I)}{P(B \mid I)}.$$

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

$$P(A, B \mid I) = P(B \mid I) \times P(A \mid B; I) = P(A \mid I) \times P(B \mid A; I),$$

which we recognize as **Bayes' Rule**.

Classical Example: a family has two puppies that are not twins. What is the probability that the youngest puppy is female given that at least one of the puppies is female?[6]

Solution: our answer to this question follows a frequentist approach – we generate trials and identify successful events. There are 4 possibilities:

$$\{\mathrm{MM}, \mathrm{MF}, \mathrm{FM}, \mathrm{FF}\}.$$

Let A and B be the events that the youngest puppy is female and that at least one puppy is female, respectively; then

$$A \mid I = \{\mathrm{FF}, \mathrm{MF}\} \quad \text{and} \quad B \mid I = \{\mathrm{FF}, \mathrm{MF}, \mathrm{FM}\},$$
$$\Longrightarrow \; P(A \mid B; I) = \frac{P(A \cap B \mid I)}{P(B \mid I)} = \frac{2/4}{3/4} = 2/3.$$

26.1.2 Bayes' Theorem

Bayes' Theorem provides an expression for the conditional probability of A given B, that is:

$$\begin{aligned} P(A \mid B; I) &= \frac{P(B \mid A; I) \times P(A \mid I)}{P(B \mid I)} \\ &= \frac{P(B \mid A; I) \times P(A \mid I)}{P(B \mid A; I) \times P(A \mid I) + P(B \mid -A; I) \times P(-A \mid I)}, \end{aligned}$$

which is a direct application of the **Law of Total Probability**.

Bayes' Theorem can be thought of as a way of **coherently updating our uncertainty in the light of new evidence**. The use of a probability distribution as a 'language' to express our uncertainty is not an arbitrary

[5] Which is to say, $\sum_k P(Y_k \mid I) = 1$ and $P(Y_j, Y_k \mid I) = 0$ for all $j \neq k$).

[6] Assume that male and female puppies are equally likely to be born.

choice: it can in fact be determined from deeper principles of logical reasoning or rational behaviour.

Example: consider a medical clinic (in what follows, we drop the explicit dependence on I to lighten the notation, but it is important to remember that it is there nonetheless).

- A could represent the event "Patient has liver disease." Past data suggests that 10% of patients entering the clinic have liver disease: $P(A) = 0.10$.
- B could represent the litmus test "Patient is alcoholic." Perhaps 5% of the clinic's patients are alcoholics: $P(B) = 0.05$.
- $B \mid A$ could represent the scenario that a patient is alcoholic, given that they have liver disease: perhaps we have $P(B \mid A) = 0.07$, say.

According to Bayes' Theorem, then, the probability that a patient has liver disease assuming that they are alcoholic is

$$P(A \mid B) = \frac{0.07 \times 0.10}{0.05} = 0.14$$

While this is a (large) increase over the original 10% suggested by past data, it remains unlikely that any particular patient has liver disease.

Bayes' Theorem with Multiple Events Let D represent some observed data and let A, B, and C be mutually exclusive (and exhaustive) events conditional on D. Note that

$$\begin{aligned} P(D) &= P(A \cap D) + P(B \cap D) + P(C \cap D) \\ &= P(D \mid A)P(A) + P(D \mid B)P(B) + P(D \mid C)P(C). \end{aligned}$$

According to Bayes' theorem,

$$\begin{aligned} P(A \mid D) &= \frac{P(D \mid A)P(A)}{P(D)} \\ &= \frac{P(D \mid A)P(A)}{P(D \mid A)P(A) + P(D \mid B)P(B) + P(D \mid C)P(C)}. \end{aligned}$$

In general, if there are n exhaustive and mutually exclusive outcomes $A_1, \dots, A_n$, we have, for any $i \in \{1, \dots, n\}$:

$$P(A_i \mid D) = \frac{P(A_i)P(D \mid A_i)}{\sum_{k=1}^{n} P(A_k)P(D \mid A_k)}$$

The denominator is simply $P(D)$, the **marginal distribution** of the data.

Note that, if the values of A_i are portions of the continuous real line, the sum may be replaced by an integral.

Example: In the 1996 General Social Survey, for males (age 30+):

- 11% of those in the lowest income quartile were college graduates.
- 19% of those in the second-lowest income quartile were college graduates.
- 31% of those in the third-lowest income quartile were college graduates.
- 53% of those in the highest income quartile were college graduates.

What is the probability that a college graduate falls in the lowest income quartile?

Solution: let Q_i represent the income quartiles ($P(Q_i) = 0.25$) and D represent the event that a male over 30 is a college graduate. Then

$$P(Q_1 \mid D) = \frac{P(D \mid Q_1)P(Q_1)}{\sum_{k=1}^{4} P(Q_k)P(D \mid Q_k)} = \frac{(0.11)(0.25)}{(0.11+0.19+0.31+0.53)(0.25)} = 0.09.$$

26.1.3 Bayesian Inference Basics

Bayesian statistical methods start with existing prior beliefs, and update these using data to provide posterior beliefs, which may be used as the basis for inferential decisions:

$$\underbrace{P(\boldsymbol{\theta} \mid D)}_{\text{posterior}} = \underbrace{P(\boldsymbol{\theta})}_{\text{prior}} \times \underbrace{P(D \mid \boldsymbol{\theta})}_{\text{likelihood}} / \underbrace{P(D)}_{\text{evidence}},$$

where the evidence is

$$P(D) = \int P(D \mid \boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta} \quad \text{or} \quad P(D) = \sum_k P(D \mid A_k)P(A_k),$$

where $\{A_k\}$ is mutually exclusive and exhaustive.

In the vernacular of Bayesian data analysis (BDA),

- the **prior**, $P(\boldsymbol{\theta})$, represents the strength of the belief in $\boldsymbol{\theta}$ without taking the observed data D into account;
- the **posterior**, $P(\boldsymbol{\theta} \mid D)$, represents the strength of our belief in $\boldsymbol{\theta}$ when the observed data D is taken into account;
- the **likelihood**, $P(D \mid \boldsymbol{\theta})$, is the probability that the observed data D would be generated by the model with parameter values $\boldsymbol{\theta}$, and
- the **evidence**, $P(D)$, is the probability of observing the data D according to the model, determined by summing (or integrating) across all possible parameter values and weighted by the strength of belief in those parameter values.

Central Data Analysis Question Bayes' Theorem allows is an essential component of the **scientific method** and knowledge discovery in general. Indeed, assume that an experiment has been conducted to determine the degree of validity of a particular hypothesis, and that corresponding experimental data has been collected.

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

Given everything that was known prior to the experiment, does the collected/observed data support (or invalidate) the hypothesis/presence of a certain condition?

The **problem** is that this is usually impossible to compute directly. Bayes' Theorem offers a **possible solution**:

$$
\begin{aligned}
P(\text{hypothesis} \mid \text{data}; I) &= \frac{P(\text{data} \mid \text{hypothesis}; I) \times P(\text{hypothesis} \mid I)}{P(\text{data} \mid I)} \\
&\propto P(\text{data} \mid \text{hypothesis}; I) \times P(\text{hypothesis} \mid I);
\end{aligned}
$$

the hope is that the terms on the right might be easier to compute than those on the left:

- $P(\text{hypothesis} \mid I)$ is the degree-of-belief that the hypothesis is true, **prior to the experiment**;
- $P(\text{hypothesis} \mid \text{data}; I)$ is the degree-of-belief that the hypothesis is true, **after the experimental data is taken into account**;
- $P(\text{data} \mid \text{hypothesis}; I)$ is the probability of observing experimental data, **assuming that the hypothesis is true**, and
- $P(\text{data} \mid I)$ is the probability of the experimental data being observed, **independently of the hypothesis**.

The theorem is often presented as

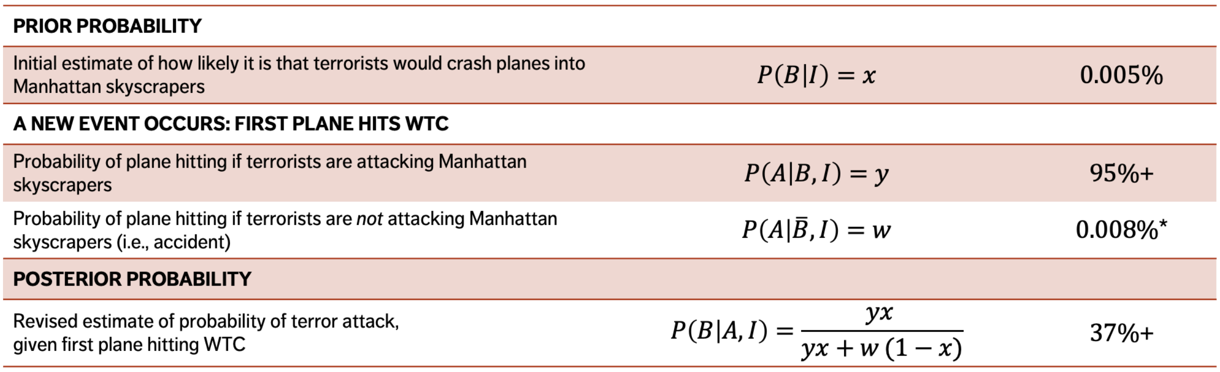
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \propto \text{likelihood} \times \text{prior},$$

i.e., **beliefs should be updated in the presence of new information**.

Example: "Most of us would have assigned almost no probability to terrorists crashing planes into buildings in Manhattan when we woke up on 9/11. 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." [323]

Let A represent the proposition that a plane crashes into Manhattan skyscrapers. Let B represent the proposition that terrorists would attack Manhattan skyscrapers; before 2001, most people would only have assigned a miniscule probability to such an event, say 0.005%. There had been two incidents of planes crashing into Manhattan skyscrapers in the previous 25,000 days before September 11, 2001, so we might assign $P(A \mid -B; I) = 0.008\%$.

We could also assign a fairly high probability of a plane hitting a Manhattan skyscraper if terrorists were attacking said skyscrapers, say $P(A \mid B; I) = 95\%$.



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%+

After one plane hitting the World Trade Center, our revised estimate of the probability of a terror attack now stands at roughly 37. If a second plane hits the World Trade Center shortly after the first one, the posterior probability of a terror attack now jumps to a whopping 99.99%.

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: SECOND 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	$P(B\|A,I) = \dfrac{yx^{\#}}{yx^{\#} + w\,(1 - x^{\#})}$	99.99%+

Determining an appropriate **prior** is a source of considerable controversy. Conservative estimates (**uninformative priors**) often lead to reasonable results, but in the absence of relevant information, it might be preferable to use **maximum entropy priors** (see Section 26.3).

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, or that model was bad, so we dare not trust our computation.[7]

26.1.4 Bayesian Data Analysis

The main characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis. The process of **Bayesian data analysis** (BDA) can be idealized by dividing it into the following 3 steps:

1. Setting up a full probability model (the **prior**) – a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process (when available).
2. Conditioning on observed data (**new data**) – calculating and interpreting the appropriate posterior distribution (i.e., the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data).
3. Evaluating the fit of the model and the implications of the resulting posterior distribution (the **posterior**) – how well does the model fit the data? Are the substantive conclusions reasonable? How sensitive are the results to the modeling assumptions made in step 1? Depending on the responses, one can alter or expand the model and repeat the 3 steps.

The essence of Bayesian methods consists in identifying the **prior beliefs** about what results are likely, and then updating those according to the **collected data**.

[7] Thankfully, the evidence is rarely required on problems of **parameter estimation**: prior to the experiment, there are numerous competing hypotheses; while the priors and likelihoods will differ, the evidence will not, so it is not needed to differentiate the various hypotheses.

For example, if the current success rate of a gambling strategy is 5%, we may say that it's reasonably likely that a small strategy modification could further improve that rate by 5 percentage points, but that it is most likely that the change will have little effect, and that it is entirely unlikely that the success rate would shoot up to 30%.[8]

As the data comes in, we **update our beliefs**. If the incoming data points to an improvement in the success rate, we move our prior estimate of the effect upwards; the more data we collect, the more confident we are in the estimate of the effect and the further we can leave the prior behind.

The end result is called the **posterior** – a probability distribution describing the likely effect of the strategy.

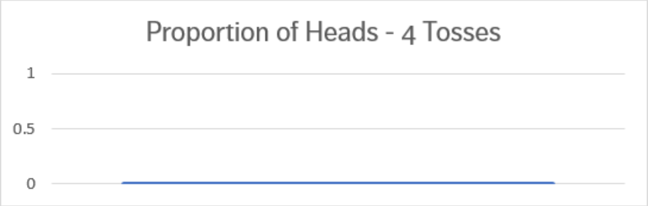
26.2 Simple Examples

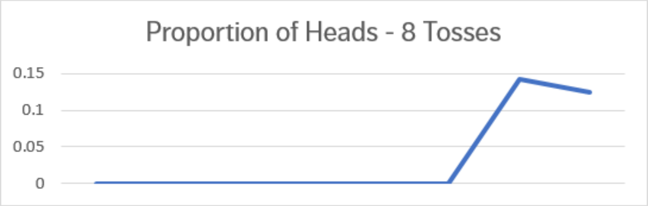
We take a look at three scenarios that will shed some light on the whole Bayesian entreprise:[9]

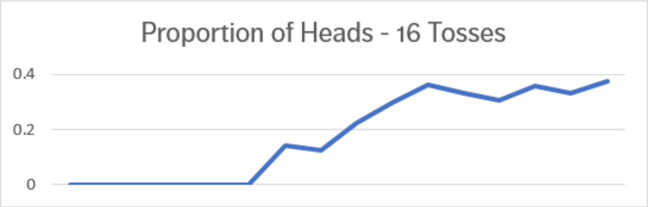
- determining if a coin is fair (or not),
- finding a link between demographic information and salary, and
- estimating the number of dollar bills in circulation.

26.2.1 The Mysterious Coin

A mysterious stranger brings back a souvenir coin from a trip to a strange and distant land. They have been flipping it pretty much non-stop since their return. You can see the proportion of heads they obtained for 4, 8, and 16 tosses.







It might seem at first that the coin might be biased, but the proportion of heads seems to inch its way towards 50%. What is truly going on?

[8] After all, it is only a small modification.

[9] These examples will showcase how priors, likelihood, and posteriors interact.

Priors Perhaps the coin is not fair, coming as it does from a strange and distant land. Let us 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; $H \approx 0, 1$: highly biased coins). A **prior** for this scenario is a **probability density function** (p.d.f.)

$$P(\text{bias} = H) = P(H \mid I).$$

Four such priors are shown below.

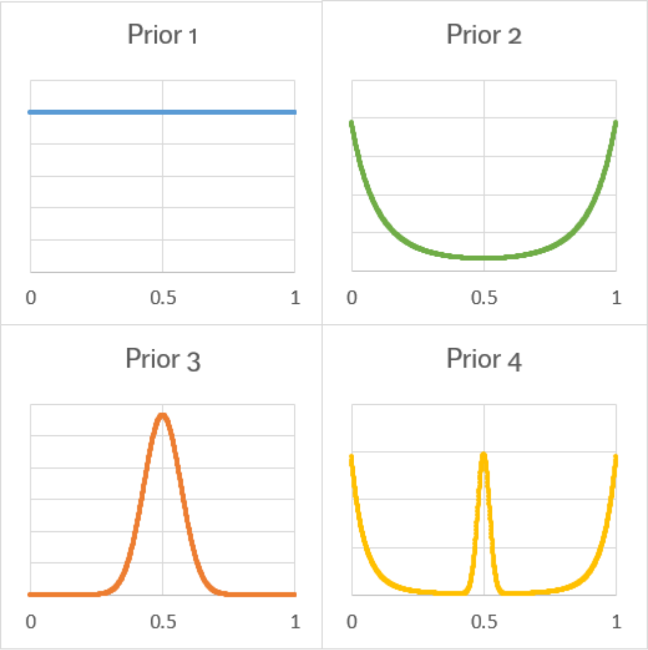


Figure 26.3: 4 priors for the fair coin problem: no idea (top left); suspect foul play (top right); just a regular coin (bottom left); probably just a regular coin, but the fact that somebody is even talking about this is suspicious (bottom right).

Why are we working with functions for the prior, when in the previous example (9/11 attacks), we only provided a number, $P(B \mid I) = 0.005\%$? In fact, we provided a **(discrete) function** as a prior:

$$P(B = x \mid I) = \begin{cases} 0.005\% & \text{if } x = \text{TRUE} \\ 99.995\% & \text{if } x = \text{FALSE} \end{cases}$$

Likelihood Let us 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:

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

The **likelihood** is the probability of observing K heads in N tosses if the bias is 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

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

Posteriors Combining the prior and the likelihood, we get:

$$P(\text{bias} = H \mid K \text{ heads}, N \text{ tosses}; I) \propto H^K(1-H)^{N-K} \times P_i(\text{bias} = H \mid I),$$

where i indexes the various prior scenarios described above.

We should thus be able to estimate the bias H^* by studying the posterior distribution for each of the 4 priors, for various number of throws N (see Figure 26.4):

- with the **non-committal prior** (blue p.d.f.)
  $$P_1(\text{bias} = H \mid I) \propto 1,$$
  the posterior is simply proportional to the likelihood; the central limit theorem seems to kick in after ≈ 30 tosses;
- with the **foul play prior** (green p.d.f.), we suspect early on that the bias is smaller than 0.5; the subsequent series of tosses moves the bias to a value $0.25 \leq H^* \leq 0.40$ quickly, as was the case with the non-informative prior – note the shrinking of the posterior with an increasing number of tosses;
- with the **regular coin prior** (orange p.d.f.)
  $$P_1(\text{bias} = H \mid I) \sim \mathcal{N}(0.5, \sigma^2),$$
  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 $0.25 \leq H^* \leq 0.40$), but the series of tosses forces the posterior to a biased distribution (note the smoother convergence of the posterior;
- with the **doubtful prior** (yellow p.d.f.), the competing hypotheses compete before converging to a bias, again in $0.25 \leq H^* \leq 0.40$. The convergence is more haphazard: as soon as one head or one tail is observed, the process nixes the two-sided coin option. Note the slower (and weirder) convergence to a gaussian posterior.

In the fair coin example, it would seem that the choice of a prior does not have much of an effect on the posterior . . . **given enough data**.

This will not always be the case.

26.2.2 The Salary Question

Income information has been collected for 4782 individuals, together with demographic details: self-reported gender, age group, and education level (1 for post-secondary degree, 0 otherwise). The table below shows some of the summary statistics for the dataset; the dataset is available in `Salary.xlsx`.

Question: is there a link between demographic information and income? How would we answer this question using classical statistical methods? What if we had reason to suspect that reported incomes follow a (potentially) different distribution for each group? Would that change the approach?

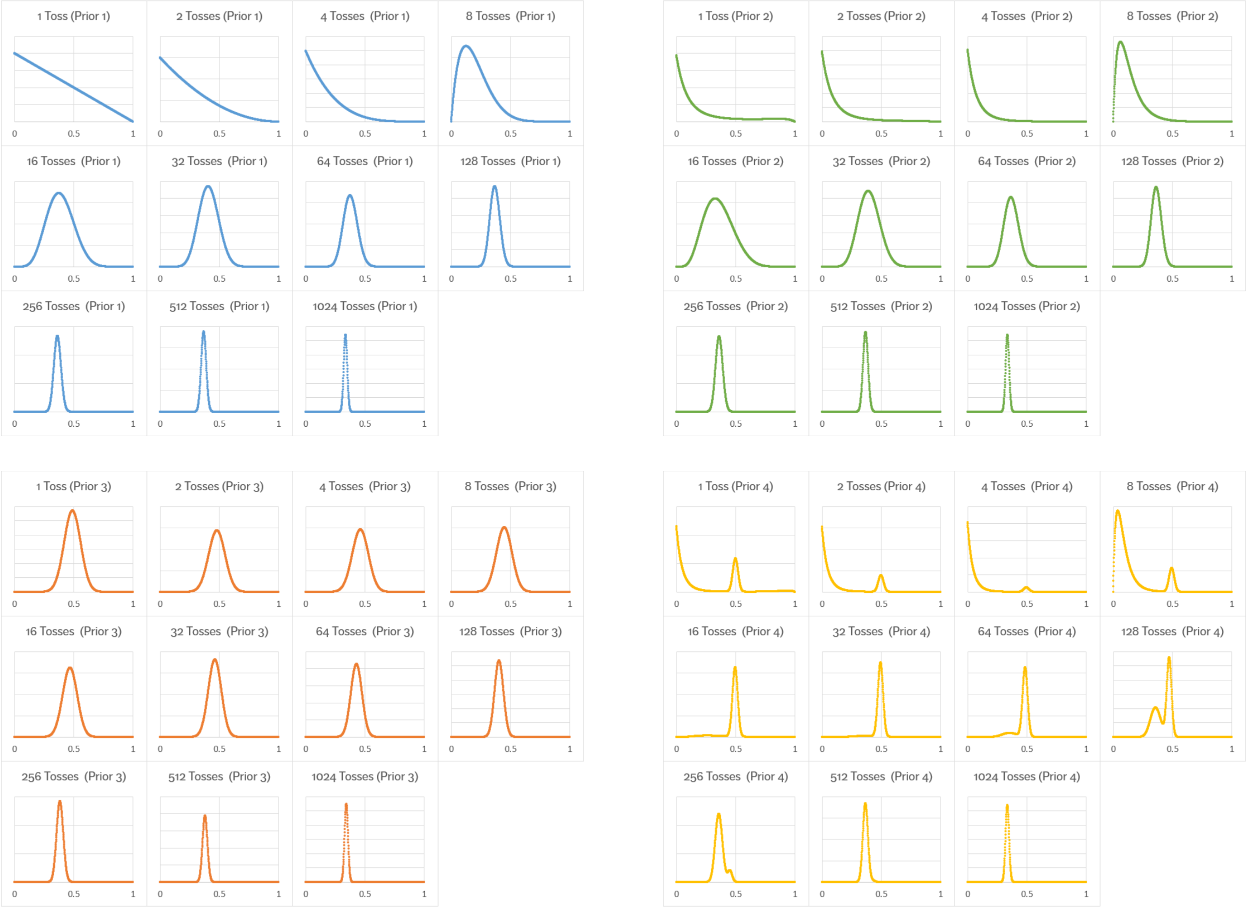
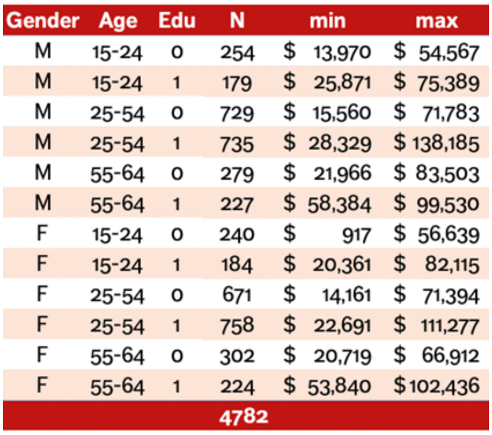


Figure 26.4: Posteriors for a different numbers of tosses; 4 priors, same data. After 128 tosses starting with the non-committal prior, we are fairly certain that the coin must be biased, with $0.25 \leq H^* \leq 0.40$ (top left); it takes roughly 256 tosses starting with the foul play prior for the same notion to arise (top right); after 512 tosses starting with the regular coin prior, we are fairly certain that the coin must be biased, with $0.33 \leq H^* \leq 0.40$ (bottom left); which is more or less the same when starting with the doubtful prior (bottom right).



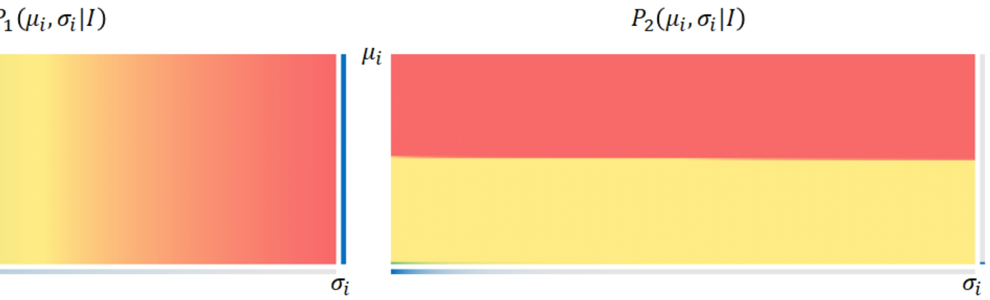
Gender	Age	Edu	N	min	max
M	15-24	0	254	$ 13,970	$ 54,567
M	15-24	1	179	$ 25,871	$ 75,389
M	25-54	0	729	$ 15,560	$ 71,783
M	25-54	1	735	$ 28,329	$ 138,185
M	55-64	0	279	$ 21,966	$ 83,503
M	55-64	1	227	$ 58,384	$ 99,530
F	15-24	0	240	$ 917	$ 56,639
F	15-24	1	184	$ 20,361	$ 82,115
F	25-54	0	671	$ 14,161	$ 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		

In the Bayesian framework, we are interested in the posterior distribution

$$P(\text{parameters} \mid \text{data}; i, I), \quad i = 1, \ldots, 12.$$

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

$$P(\mu_i, \sigma_i \mid \text{reported salaries in group } i; I), \quad i = 1, \ldots, 12.$$



r the salary problem. Green represents large probabilities; red, low probabilities. The blue zones represent
gher values.

Priors Determining a reasonable collection of **priors**

$$P(\mu_i, \sigma_i \mid I), \quad \text{for } i = 1, \ldots, 12,$$

is no easy task. One could naively pick a joint distribution which "**peaks**" at the sample mean $\overline{x}_i$, with standard deviation s_i, for each 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" [37, 324]. This translates into using a **non-informative prior**

$$P_1(\mu_i, \sigma_i \mid I) \propto \sigma_i^{-1}, \quad i = 1, \ldots, 12$$

(we will discuss these further in the next section).

For comparison's sake, we will also consider the prior

$$P_2(\mu_i, \sigma_i \mid I) \propto \mu_i^{500} \sigma_i^{-4}, \quad i = 1, \ldots, 12.$$

The two priors are illustrated in Figure 26.5.

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.[10]

Instead, we could only consider them over a suitable finite sub-region; or use the fact that the product of the likelihood and the prior *is* normalizable.

Likelihood Let us denote the number of observations in group i by N_i. The likelihood is the probability

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

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

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

[10] The integral of these priors over the positive quadrant is infinite.

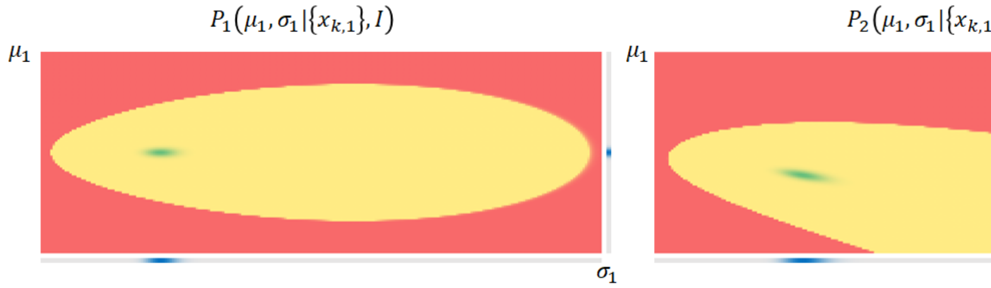
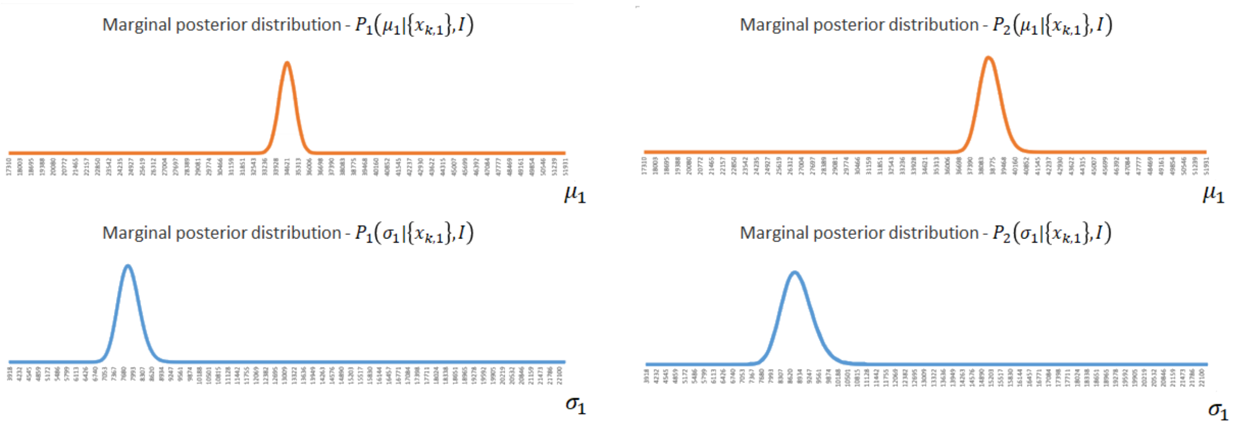

$P_1(\mu_1,\sigma_1|\{x_{k,1}\},I)$ $P_2(\mu_1,\sigma_1|\{x_{k,1}$

Figure 26.6: Posteriors for the salary problem (one per prior), for group $i = 1$. Green represents large probabil
Note the shape of the posteriors. The blue zones represent marginal probabilities of higher values.


Marginal posterior distribution - $P_1(\mu_1|\{x_{k,1}\},I)$

Marginal posterior distribution - $P_2(\mu_1|\{x_{k,1}\},I)$

Marginal posterior distribution - $P_1(\sigma_1|\{x_{k,1}\},I)$

Marginal posterior distribution - $P_2(\sigma_1|\{x_{k,1}\},I)$

Figure 26.7: Marginal posteriors for the salary problem (one for each of the priors), for group $i = 1$. Note the differences in the distributions for each scenario.

Posteriors Combining the prior and the likelihood, we get, for the first prior:

$$P_1(\mu_i,\sigma_i \mid \{x_{k,i}\};I)$$
$$\propto \sigma_i^{-(N_i+1)} \prod_{k=1}^{N_i} \exp\left(\frac{-(\mu_i - x_{k,i})^2}{2\sigma_i^2}\right), \quad i = 1,\dots,12,$$

while for the second prior:

$$P_2(\mu_i,\sigma_i \mid \{x_{k,i}\};I)$$
$$\propto \mu_i^{500}\sigma_i^{-(N_i+4)} \prod_{k=1}^{N_i} \exp\left(\frac{-(\mu_i - x_{k,i})^2}{2\sigma_i^2}\right), \quad i = 1,\dots,12,$$

over some suitable sub-region in parameter space.

The joint posterior distributions for (μ_1,σ_1) (one for each of the priors) when $i = 1$ are shown in Figure 26.6.

We can read the likely values of each parameters for each scenario by looking at the spikes in the marginal posteriors of Figure 26.7.

The first two examples were (somehow ashamedly) conducted with Excel; the next example shows how we can use programmatical tools (like R) to answer questions using Bayesian analysis.

26.2.3 Money ($ Bill Y'All)

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

The **problem:** we cannot count them all – so what do we do?

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

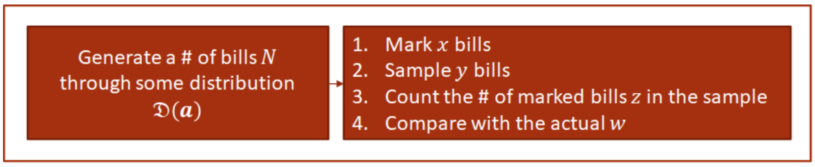
1. Capture a few 5$ bills.
2. Mark them and put them back in circulation.
3. At some later point, capture a few 5$ bills.
4. Count how many are marked.

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

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

Unlike the previous examples where we were trying to estimate the parameters from the data, we are trying to estimate data from parameters (**generative model**) – we do not compute the likelihood directly.

Simple Model In the simplest model, we might proceed as follows:



Generate a # of bills N through some distribution $\mathfrak{D}(\boldsymbol{a})$

1. Mark x bills
2. Sample y bills
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, w are given; z, N to be found

ase schematics

1. We start by drawing a **large** random sample of # of bills N from an acceptable "prior" distribution on the parameters.
2. Using the Ns and the generative model (with x and y given – the observed values), we produce a (synthetic) # of marked bills z in each sample.
3. Finally, we only retain those values of N for which $z = w$.

Let us implement this in R using the values of x, y, and w provided above. We will generate priors using 500,000 replicates:

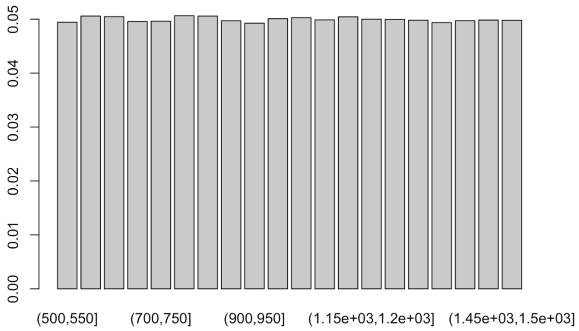
```
set.seed(1) # for replicability
N.draw = 500000    # number of replicates
x = 500 # number of bills marked in the initial capture
y = 300 # number of bills sampled in the second capture
w = 127 # number of marked bills in the second capture
```

Since $x = 500$ were first "captured", we know that there are at least 500 bills in circulation. To keep things from getting out of hands, we select a theoretical maximum for the number of bills in circulation.

```
upper.limit = 1500 # maximum (theoretical) number of bills
bin.width = 50     # for plotting the posterior
```

We now draw to create the prior distribution on the possible number of bills N_{bills} in circulation:

```
N.bills = sample (x:1500, N.draw, replace=TRUE)
barplot(table(cut(N.bills, seq(x, upper.limit, bin.width))) /
          length(N.bills), col = "gray")
```



A priori, all of these are "equally likely". Now, we use the observed "catch-and-release" data to define the generative model, in which we capture $x = 500$ bills in the first round, and $y = 300$ in the second:

```
pick.bills <- function(N.bills) {
  bills <- rep(0:1, c(N.bills - x, x)) # 0 for un-marked
                     # 1 for marked in the inital capture
  sum(sample(bills, y)) # sampling y bills in the 2nd round
}
```

The number of re-captured bills (for each trial) is simulated below:

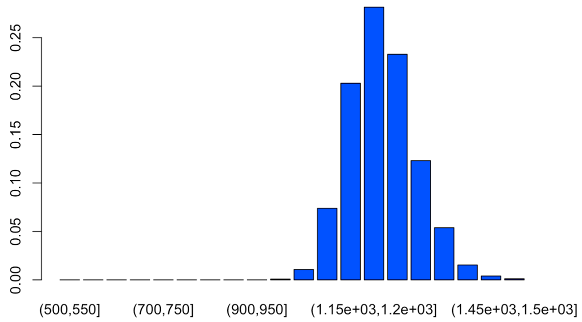
```
N.marked <- rep(NA, N.draw)
  for(i in 1:N.draw) {
    N.marked[i] <- pick.bills(N.bills[i])
  }
```

In the language of the generative model, `N.marked` is z. Now, we only keep those trials for which there were $w = 127$ re-captured marked bills, and retain the number of bills in circulation for these trials:

```
post.bills <- N.bills[N.marked == w]
```

Finally, we plot the posterior distribution:

```
barplot(table(cut(post.bills, seq(x,upper.limit,bin.width))) /
          length(post.bills), col = "blue")
```

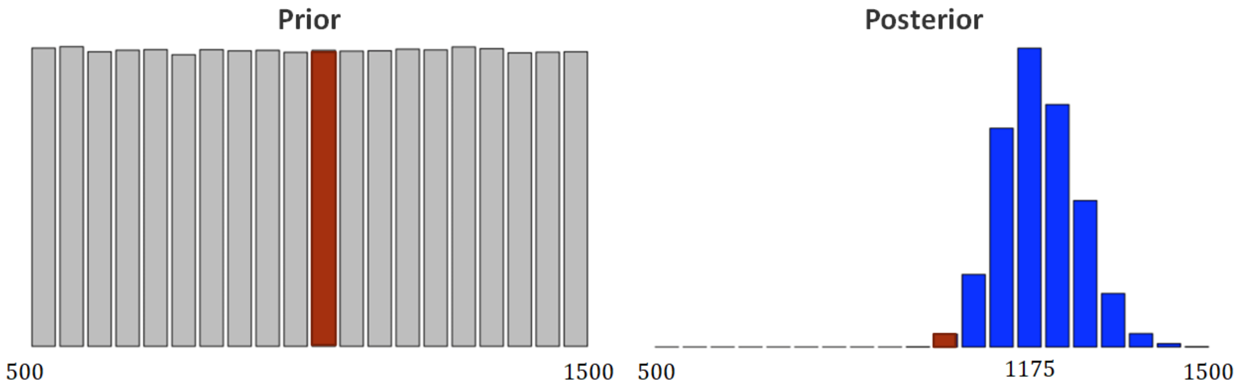


The summary statistics for the posterior distribution of the number of bills in circulation is thus:

```
length(post.bills)
summary(post.bills)
```

```
[1] 4754
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    979    1143    1188    1193    1236    1492
```

In other words, out of 500,000 trials, a little fewer than 5000 had the right characteristics (x, y, and w as observed in the "real world"), and the average/median number of bills in circulations for this smaller subset of trials is a tad below 1200. The Bayesian situation is illustrated below.[11]



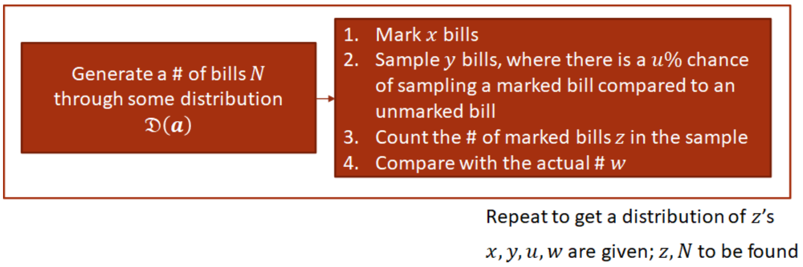
$$P(N = 1000|z = 127, I) \propto P(z = 127|N = 1000, I) \times P(N = 1000|I)$$

Model: 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 probability $u = 90\%$, say).

In this case, we might proceed as follows:

1. We start by drawing a **large** random sample of # of bills N from an acceptable "prior" distribution on the parameters.
2. Using the Ns and the generative model (with x, y, and u given – the observed values), we produce a (synthetic) # of marked bills z in each sample.
3. Finally, we only retain those values of N for which $z = w$.

[11] We used a different seed, so the charts are slightly different, but the main ideas



Generate a # of bills N through some distribution $\mathfrak{D}(\boldsymbol{a})$

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

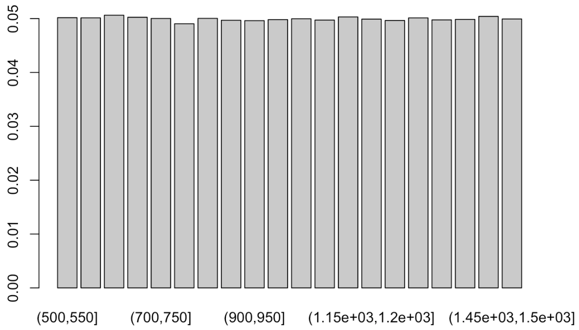
Repeat to get a distribution of z's

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

Figure 26.9: C
in the brittle r

Let us implement this in R using the values of x, y, u, and w provided above. We will generate priors using 500,000 replicates:

```
set.seed(10) # for replicability
N.draw = 500000    # number of replicates
x = 500 # number of bills marked in the initial capture
y = 300 # number of bills sampled in the second capture
w = 127 # number of marked bills in the second capture
u = 0.9 # probability that marked bills will be retired
upper.limit = 1500 # maximum (theoretical) number of bills
bin.width = 50     # for plotting the posterior
N.bills = sample (x:1500, N.draw, replace=TRUE)
barplot(table(cut(N.bills, seq(x, upper.limit, bin.width))) /
          length(N.bills), col = "gray")
```



A priori, all of these are "equally likely" in the brittle scenario too. Now, we use the observed "catch-and-release" data to define the generative model, in which we capture $x = 500$ bills in the first round, and $y = 300$ in the second round, knowing that $u = 0.9$ of first round marked bills will be retired.[12]

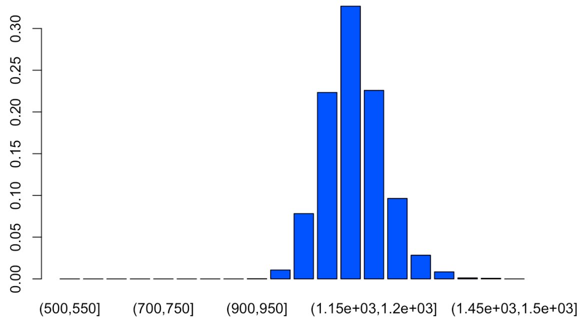
```
pick.bills <- function(N.bills) {
  bills <- rep(0:1, c(N.bills - x, x))
  prob.pick <- ifelse(bills == 0, 1.0, u) # brittleness
  sum(sample(bills, y, prob = prob.pick))
}
```

The number of re-captured bills (for each trial) is simulated below:

```
N.marked <- rep(NA, N.draw)
  for(i in 1:N.draw) {
```

[12] Would we expect there to be more bills in circulation, given these observations, in the brittle case or the simple case?

```
    N.marked[i] <- pick.bills(N.bills[i])
  }

# Posterior distribution
post.bills <- N.bills[N.marked == w]
barplot(table(cut(post.bills, seq(x,upper.limit,bin.width))) /
          length(post.bills), col = "blue")
```



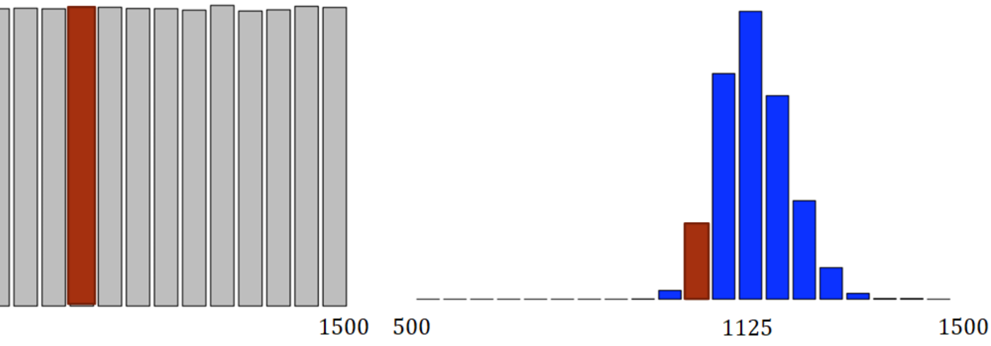
The summary statistics for the posterior distribution of the number of bills in circulation is thus:

```
length(post.bills)
summary(post.bills)
```

```
[1] 4410
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    935    1089    1129    1132    1172    1411
```

In other words, out of 500,000 trials, about 4400 had the right characteristics (x, y, u, and w as observed in the "real world"), and the average/median number of bills in circulations for this smaller subset of trials is a roughly 1130. Does this make sense, given the brittleness assumption?

The Bayesian situation is illustrated below.[13]



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

[13] Again, the charts are slightly different due the use of a different seed.

Model: Listen to the Banker Let us say that an old banker thinks that there should be about 1000 bills in circulation. How can we incorporate this piece of information?

In this case, we might proceed as follows:

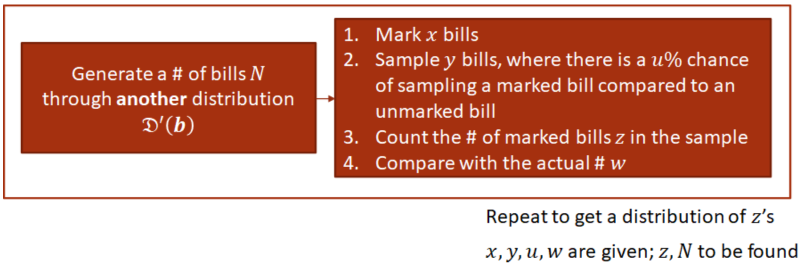

Generate a # of bills N through **another** distribution $\mathfrak{D}'(\boldsymbol{b})$
1. Mark x bills
2. Sample y bills, where there is a $u\%$ chance of sampling a marked bill compared to an unmarked bill
3. Count the # of marked bills z in the sample
4. Compare with the actual # w
Repeat to get a distribution of z's
x, y, u, w are given; z, N to be found

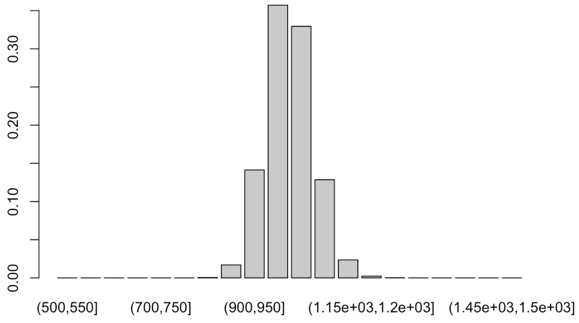
Figure 26.10:
ics in the expe

Let us implement this in R using the values of x, y, u, and w provided above, as well as the expert's best guess. We will generate priors using 500,000 replicates:

```
set.seed(100) # for replicability
N.draw = 500000 # number of replicates
x = 500 # number of bills marked in the initial capture
y = 300 # number of bills sampled in the second capture
w = 127 # number of marked bills in the sample
u = 0.9 # probability that marked bills will be retired
banker.mean = 1000 # banker guess
upper.limit = 1500 # maximum (theoretical) number of bills
bin.width = 50     # for plotting the posterior
```

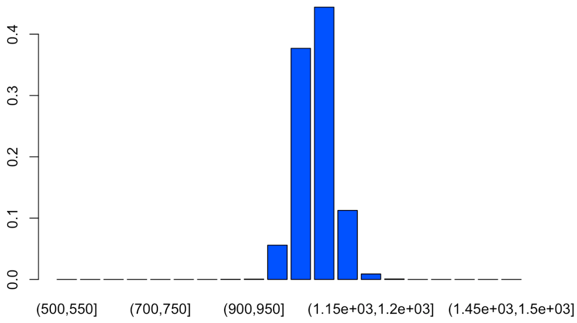
We now draw to create the prior distribution on the possible number of bills N_{bills} in circulation, using the banker's experience (instead of a uniform distribution, the prior might follow a binomial distribution with mean 1000$, say).

```
N.bills = rnbinom(N.draw, mu = banker.mean - x, size = w) + x
barplot(table(cut(N.bills, seq(x, upper.limit, bin.width))) /
          length(N.bills), col = "gray")

pick.bills <- function(N.bills) {
  bills <- rep(0:1, c(N.bills - x, x))
  prob.pick <- ifelse(bills == 0, 1.0, u)
  sum(sample(bills, y, prob = prob.pick))  second capture
}
```



The number of re-captured bills (for each trial) is simulated below:

```
N.marked <- rep(NA, N.draw)
  for(i in 1:N.draw) {
    N.marked[i] <- pick.bills(N.bills[i])
  }

# Posterior
post.bills <- N.bills[N.marked == w]
barplot(table(cut(post.bills, seq(x,upper.limit,bin.width))) /
          length(post.bills), col = "blue")
```

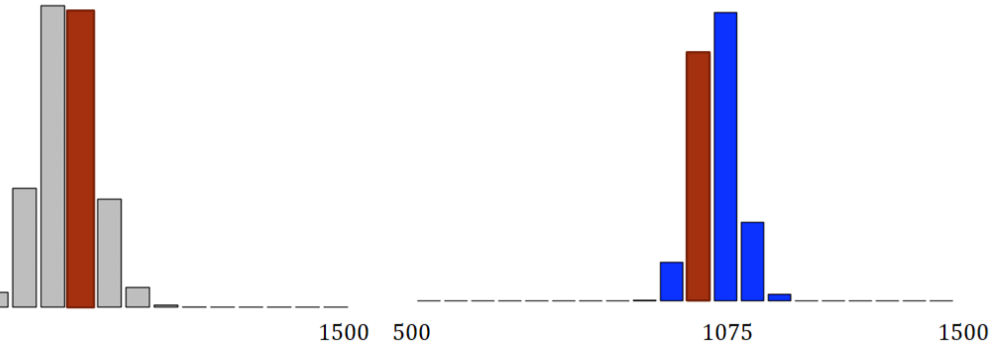


The summary statistics for the posterior distribution of the number of bills in circulation is thus:

```
length(post.bills)
summary(post.bills)
```

```
[1] 5258
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    893    1031    1057    1058    1083    1209
```

In other words, out of 500,000 trials, about 5250 had the right characteristics (x, y, u, and w as observed in the "real world"), and the average/median number of bills in circulations for this smaller subset of trials is a roughly 1050. Does this make sense, given the banker's opinion and the observations? The Bayesian situation is illustrated below.



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

26.3 Prior Distributions

Specifying a model means, by necessity, providing a prior distribution for the unknown parameters $\boldsymbol{\theta}$. The prior plays a critical role in Bayesian inference through the updating statement :

$$P(\boldsymbol{\theta} \mid D) \propto P(\boldsymbol{\theta}) \times P(D \mid \boldsymbol{\theta}).$$

In the Bayesian approach, all unknown quantities are described probabilistically, even before the data has been observed.

All priors are **subjective** in the sense that the decision to use a prior is left completely up to the researcher. But the choice of priors **is no more subjective than the choice of likelihood, the selection or collection of a sample, the estimation, or the statistic used for data reduction**. The choice of a prior can substantially affect posterior conclusions, however, especially with small sample sizes.

26.3.1 Conjugate Priors

The main challenge of Bayesian methods is that the posterior distribution of the vector $\boldsymbol{\theta}$ might not have an analytical form. Specifically, producing marginal posterior distributions from high-dimensional posteriors by repeated analytical integration may be difficult or even impossible mathematically.

There are exceptions however, providing easily obtainable computational posteriors through the use of a **conjugate prior**. Conjugacy is a joint property of a prior and a likelihood implying that the posterior has the same distributional form as the prior, but with different parameter(s).

The table below represents some common likelihoods and their conjugate priors (an extensive list can be found in [325]).

Likelihood	Prior	Hyperparameters
Bernoulli	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	$\alpha > 0, \beta > 0$

For instance, if the probability of s successes in n trials (the **likelihood**) is given by

$$P(s, n \mid q) = \frac{n!}{s!(n-s)!} q^s (1-q)^{n-s}, \quad q \in [0,1],$$

and the **prior probability** for q follows a Beta(α, β) distribution with $\alpha > 0, \beta > 0$, so that

$$P(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)}, \quad \text{for } q \in [0,1],$$

then the **posterior distribution** for q given s successes in n trials follows a Beta$(\alpha + s, \beta + n - s)$ distribution, so that

$$P(q \mid s, n) = \frac{P(s, n \mid q) \times P(q)}{P(s, n)} = \frac{q^{\alpha+s-1}(1-q)^{\beta+n-s-1}}{B(\alpha + s, \beta + n - s)}, \quad \text{for } q \in [0, 1].$$

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**.

26.3.2 Uninformative Priors

An **uninformative prior** (or **objective** prior) is one which intentionally provides very little specific information about the parameters of interest. Uninformative priors are very useful from the perspective of traditional Bayesianism seeking to mitigate the frequentist criticism of **intentional subjectivity**.

The rationale for using uninformative prior distributions is often said to be 'to let the data speak for itself,' so that inferences are unaffected by information external to the current data.

A classic uninformative prior is the **uniform prior**. A proper uniform prior integrates to a finite quantity and is thus normalizable. For example, for data following a Bernoulli(θ) distribution, a uniform prior on θ is

$$P(\theta) = 1, \quad 0 \le \theta \le 1.$$

For data with following a $N(\mu, 1)$ distribution, say,[14] the uniform prior on the support of μ is improper as

$$P(\mu) = 1, \quad -\infty < \mu < \infty$$

diverges; however, such a choice could still be acceptable as long as the resulting posterior is normalizable.[15] As there are instances where an improper prior yields an improper posterior, care is warranted.

This is also called the **principle of indifference**, which states that with no evidence one way or another, degrees of belief should be distributed equally among all the considered outcomes.[16]

There are plenty of situations where the uniform prior is **not** an appropriate prior; such a prior makes assumptions about the distribution of the parameters of interest that fall squarely in the **subjective camp**. The use of uniform priors is often justified solely on the basis of convenience.[17]

The **Jeffreys prior** is an approach to generate uninformative priors. For a given random parameter $\boldsymbol{\theta}$, the Jeffreys prior is

$$P(\boldsymbol{\theta} \mid I) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\theta})},$$

where $\mathcal{I}(\boldsymbol{\theta})$ represents the **Fisher information**, which measures the amount of information that an observable random variable X implies about an unknown parameter vector $\boldsymbol{\theta}$ (i.e., we are interested in $P(X \mid \boldsymbol{\theta})$).

[14] Or any data with **unbounded support**.

[15] Which is to say, the integral of the posterior converges on its support.

[16] But *Bertrand's Paradox* ↗ provides doubt as to the validity of this principle.

[17] Since the posterior is then simply proportional to the likelihood.

Let $f(X \mid \boldsymbol{\theta})$ be the corresponding p.d.f./p.m.f.;

$$[\mathcal{I}(\boldsymbol{\theta})]_{i,j} = -\operatorname{E}\left[\frac{\partial^2}{\partial\theta_i \, \partial\theta_j} \log f(X \mid \boldsymbol{\theta}) \,\middle|\, \boldsymbol{\theta} \right].$$

Note that the Jeffreys prior depends on underlying statitistical model:

- if X follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$, with σ fixed, and all we assume is that μ is a **location** parameter, then the Jeffrey prior would be
  $$P(\mu \mid I) \propto 1,$$
  an improper uniform distribution (all locations are equally likely to be the mean);
- if X follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$, with μ fixed, and all we assume is that $\sigma > 0$ is a **scale** parameter, then it would be
  $$P(\sigma \mid I) \propto \frac{1}{\sigma},$$
  again an improper distribution, but one for which a dispsersion σ becomes progressively less likely as it increases;
- if X follows a Poisson distribution $\mathcal{P}(\lambda)$ and all we assume is that $\lambda \geq 0$, then it would be the improper distribution
  $$P(\lambda \mid I) \propto \frac{1}{\sqrt{\lambda}}.$$

In contrast, a **weakly informative prior** is one for which only **partial information** about a variable is available; the choice of a uniform prior is often weakly informative.

We will discuss another uninformative approach, the **Maximum Entropy prior**, shortly.

26.3.3 Informative Priors

Informative priors are those that **deliberately** insert information that researchers have at hand. This seems like a reasonable approach since previous scientific knowledge should play a role in statistical inference.

However, there are two important requirements for researchers:

1. the **overt** declaration of prior specification, and
2. a 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. But where do informative priors come from, in the first place?

Generally these priors are derived from:

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

Prior information from past studies need not be in agreement. One useful strategy is to construct prior specifications from **competing school-of-thoughts** in order to contrast the resulting posteriors and produce informed statements about the relative strength of each of them.

Example: we have noted previously that a Bernoulli likelihood and a Beta prior form a set of conjugate priors. For this exercise, we use the R function `BernBeta()` defined in the excellent [326].[18]

1. Start with a prior distribution that expresses some uncertainty that a coin is fair: Beta($\theta \mid 4, 4$). Flip the coin once; assume that a Head is obtained. What is the posterior distribution of the uncertainty in the coin's fairness θ?

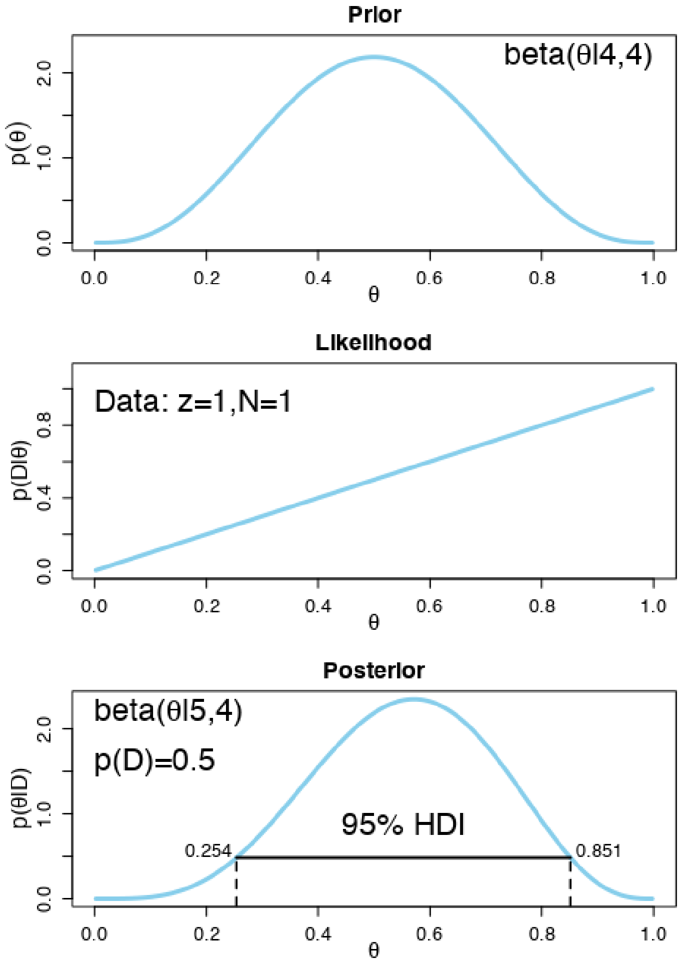
   Solution: we know, on theoretical grounds, that the posterior follows a

   $$\text{Beta}(\theta \mid 4+1, 4+1-1; I) = \text{Beta}(\theta \mid 5, 4; I)$$

   distribution.[19]

   ```
   post = BernBeta( c(4,4) , c(1) )
   show(post)
   ```

   ```
   [1] 5 4
   ```



[18] This function uses the conjugacy between the Bernoulli (likelihood) and the Beta (prior) distributions to determine the posterior distribution Beta for the uncertainty in the fairness of the coin (1 represents a H(ead) on the flip, 0 a T(ail)). Note that the function returns the posterior beta values each time it is called, so returned values can be fed back into the prior in a subsequent function call.

[19] The label on the y−axis of the posterior distribution provides the posterior parameters.

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?

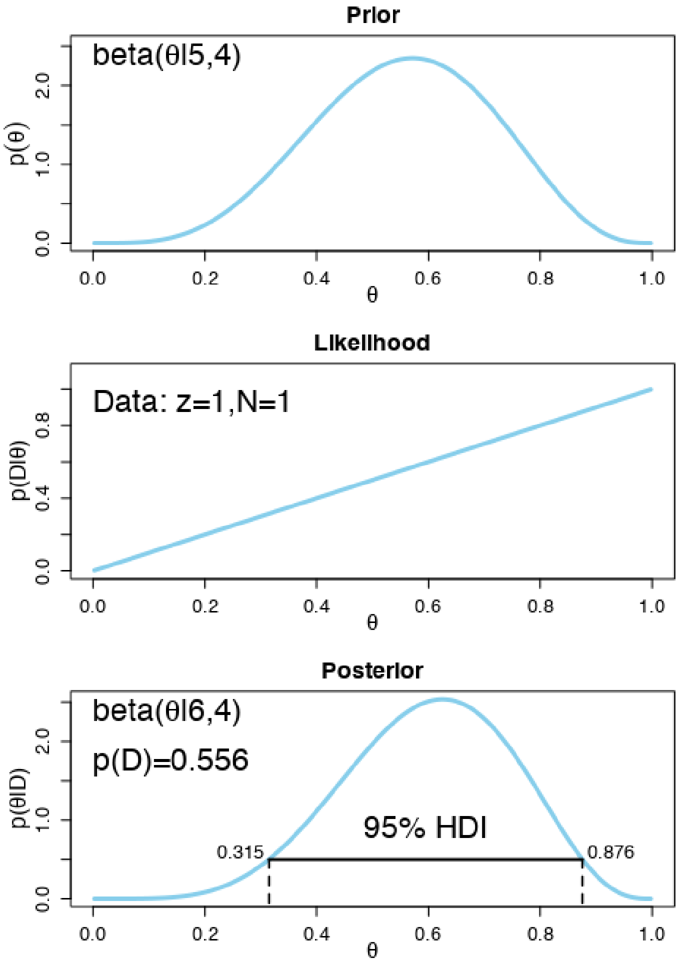
   Solution: on theoretical grounds, the posterior is

   $$\text{Beta}(\theta \mid 6, 4; I),$$

   which is shown below.

```
post = BernBeta( post , c(1) )
show(post)
```

```
[1] 6 4
```

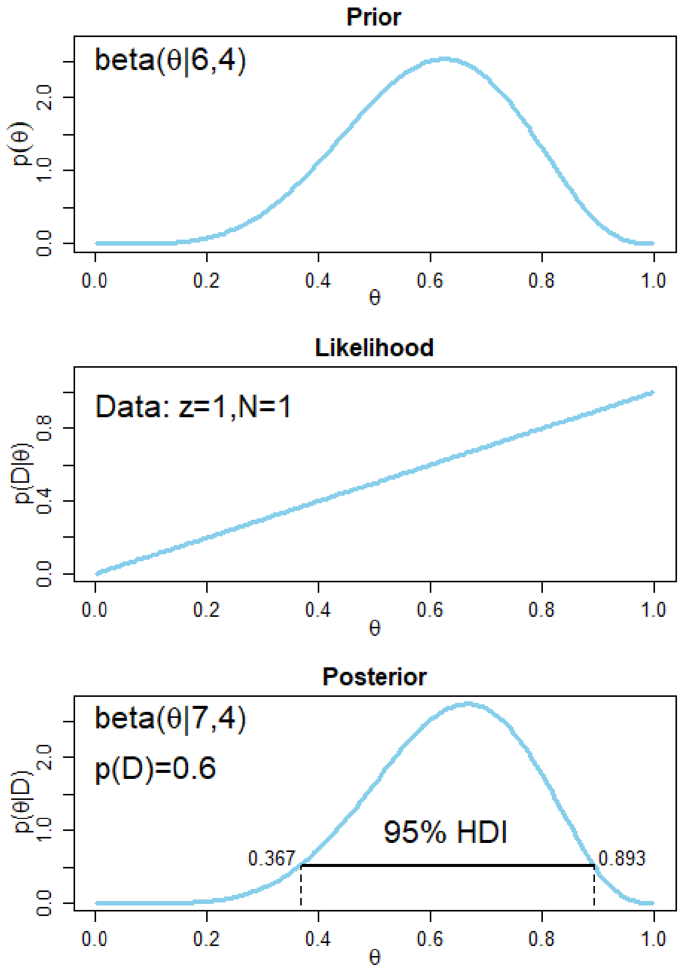


3. Using the most recent posterior as the prior for the next flip, flip a third time and obtain yet again a H. What is the new posterior?

   Solution: in this case, we know that the posterior for the coin's fairness follows a Beta$(\theta \mid 7, 4; I)$ distribution.

```
post = BernBeta( post , c(1) )
show(post)
```

```
[1] 7 4
```



> Should flipping 3 H in a row give us pause? Is there enough evidence to suggest that $\theta \neq 0.5$ (i.e, that the coin is not fair)? What if we were to flip 18 H in a row from this point on?[20]

When working on a problem, it can be easy to get side-tracked and confused with the notation. In those cases, it is useful to return to the definition of each of the terms in Bayes' theorem (i.e., $P(\theta \mid D; I)$, $P(D \mid I)$, $P(D \mid \theta; I)$, etc.).

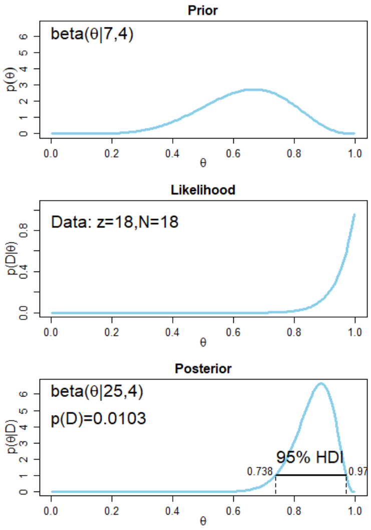
Example: suppose that a friend has a coin that we know comes from a magic store; as a result, we believe that the coin is strongly biased in either of the two directions (it could be a trick coin with both sides being H, for instance), but we don't know which one it favours. We will express the belief of this prior as a Beta distribution. Let's say that our friend flips the coin five times; resulting in 4 H and 1 T. What is the posterior distribution of the coin's fairness θ?

Solution: we start with a prior that corresponds with our assumptions, and assume 4 H and 1 T:

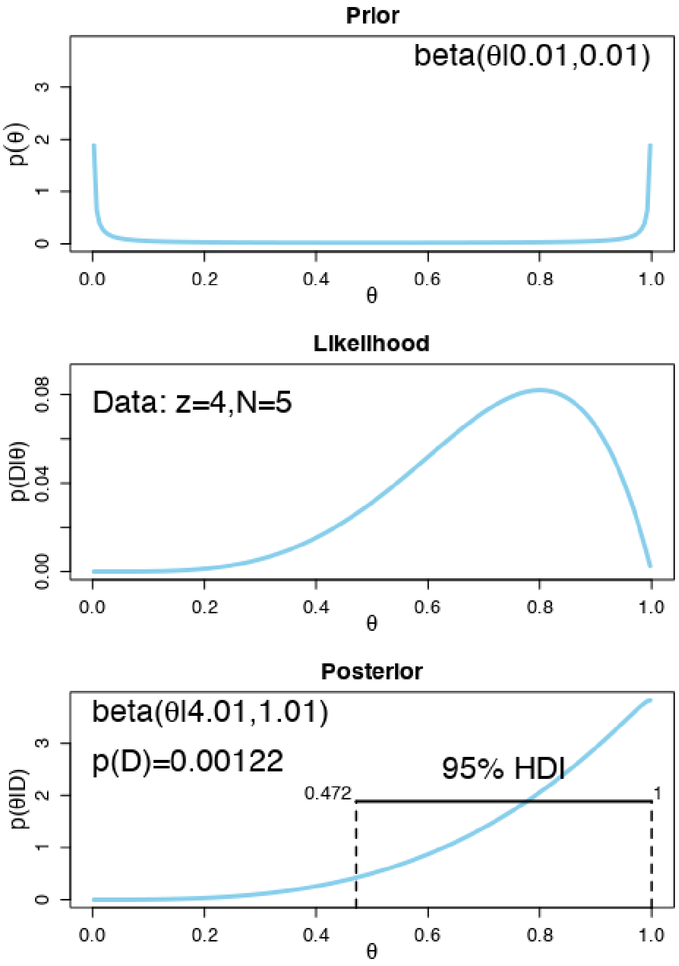
```
post = BernBeta( c(1,1)/100 , c(1,1,1,1,0) )
show(post)
```

```
[1] 4.01 1.01
```

[20] 20: The modified code would yield:



Note the posterior's bias.



This prior captures our belief that the coin is strongly biased, although we do not know in which direction the bias lies before seeing data. The use of 0.01 is arbitrary, in a sense; 0.1 would have worked as well, say.

The posterior distribution is

$$\mathrm{Beta}(\theta \mid 4.01, 1.01; I),$$

which, as shown above, has its mode essentially at 1.0, and not near the mean ≈ 0.8. Is the coin indeed biased? In which direction?

How would the answer change if we had no reason to suspect that the coin was biased in the first place? These are all questions that could be answered by playing with `BernBeta()`.

26.3.4 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.

Consider a discrete space X of cardinality M with probability density $P(X) = \mathbf{p} = (p_1, \ldots, p_M)$. The **entropy** of such a $\mathbf{p}$, denoted by $H(\mathbf{p})$, is given by

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

In the case of a continuous p.d.f. $P(\mathbf{X}) = P(X_1, \ldots, X_n)$ on some domain $\Omega \subseteq \mathbb{R}^n$, the entropy is given by

$$H(P) = -\int_{\Omega} P(\mathbf{Z}) \log(P(\mathbf{Z}))\, d\mathbf{Z}.$$

The **maximum entropy principle** (MaxEnt) states that, 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 $\mathbf{p}$ to lie in the **probability simplex**, that is, $\sum_i p_i = 1$ and $p_i \geq 0$ for all i in the discrete case, or $\int_{\Omega} P(\mathbf{Z})\, d\mathbf{Z} = 1$ and $P(\mathbf{Z}) \geq 0$ on Ω in the continuous case.

Example: without constraints, the MaxEnt principle yields a prior which solves the optimization problem:

$$\begin{array}{rl} \max & -p_1 \log p_1 - \cdots - p_M \log p_M \\ \text{s.t.} & p_1 + \cdots + p_M = 1 \text{ and } p_1, \ldots, p_M \geq 0 \end{array}$$

With the method of Lagrange multipliers, the optimization reduces to

$$\mathbf{p}^* = \arg_{\mathbf{p}} \max\{H(\mathbf{p}) - \lambda(p_1 + \cdots + p_M - 1)\},$$

whose solution is $\mathbf{p}^* \propto$ constant. Hence, subject to no additional constraints, the uniform distribution is the maximum entropy prior.

Example: use Bayesian analysis to predict the cab waiting time?

> "The joke about New York is that you can never get a cab, except when you don't need a cab, and then there are cabs everywhere" (quote and example from S.DeDeo's *Maximum Entropy Methods* tutorial [327]).

At various moments, we head out to the street to hail a cab, and we keep track of how long it took before a cab was available. Perhaps the observations (in minutes) look like this

$$6, 3, 4, 6, 2, 3, 2, 6, 4, 4.$$

What can you conclude about the waiting time for a New York cab?

Solution: in the best case scenario a cab is waiting for us as we get to the curb ($j = 0$), while in the worst case scenario (a zombie apocalypse, say?), no cab ever comes ($j \longrightarrow \infty$). But can anything else be said?

To use MaxEnt in this situation, we need to find – among all of the trial distributions that could have generated the observed waiting times – the one with the highest entropy. Unfortunately, there are infinitely many such distributions.

We can narrow the search, however, by including a constraint stating that the expected value of the trial distributions should be the same as the mean of the sample: in this case, 4.

The two constraints translate to

$$g_1(\mathbf{p}) = \sum_{j=0}^{\infty} j \cdot p_j - 4 = 0 \quad \text{and} \quad g_2(\mathbf{p}) = \sum_{j=0}^{\infty} p_j - 1 = 0,$$

where p_j is the probability of having to wait j minutes for a cab.

The method of Lagrange multipliers reduces the problem to solving

$$\arg_{\mathbf{p}} \max \{H(\mathbf{p}) - \lambda_1 g_1(\mathbf{p}) - \lambda_2 g_2(\mathbf{p})\}.$$

This requires solving the gradient equation

$$\nabla_{\mathbf{p}} H(\mathbf{p}) = \lambda_1 \nabla_{\mathbf{p}} g_1(\mathbf{p}) + \lambda_2 \nabla_{\mathbf{p}} g_2(\mathbf{p}),$$

which gives rise to equations of the form

$$-(\ln p_j + 1) = \lambda_1 j + \lambda_2, \quad j = 0, 1, \ldots,$$

or simply $p_j = \exp(-\lambda_1 j)\exp(-1 - \lambda_2)$ for $j = 0, 1, \ldots$.

Since

$$1 = \sum_{j=0}^{\infty} p_j = \exp(-1 - \lambda_2) \sum_{j=0}^{\infty} \exp(-\lambda_1 j),$$

we have

$$\exp(1 + \lambda_2) = \sum_{j=0}^{\infty} \exp(-\lambda_1 j) = \frac{1}{1 - \exp(-\lambda_1)},$$

assuming that $|\exp(-\lambda_1)| < 1$.

Similarly,

$$4 = \sum_{j=0}^{\infty} j p_j = \exp(-1 - \lambda_2) \sum_{j=0}^{\infty} j \exp(-\lambda_1 j),$$

so that

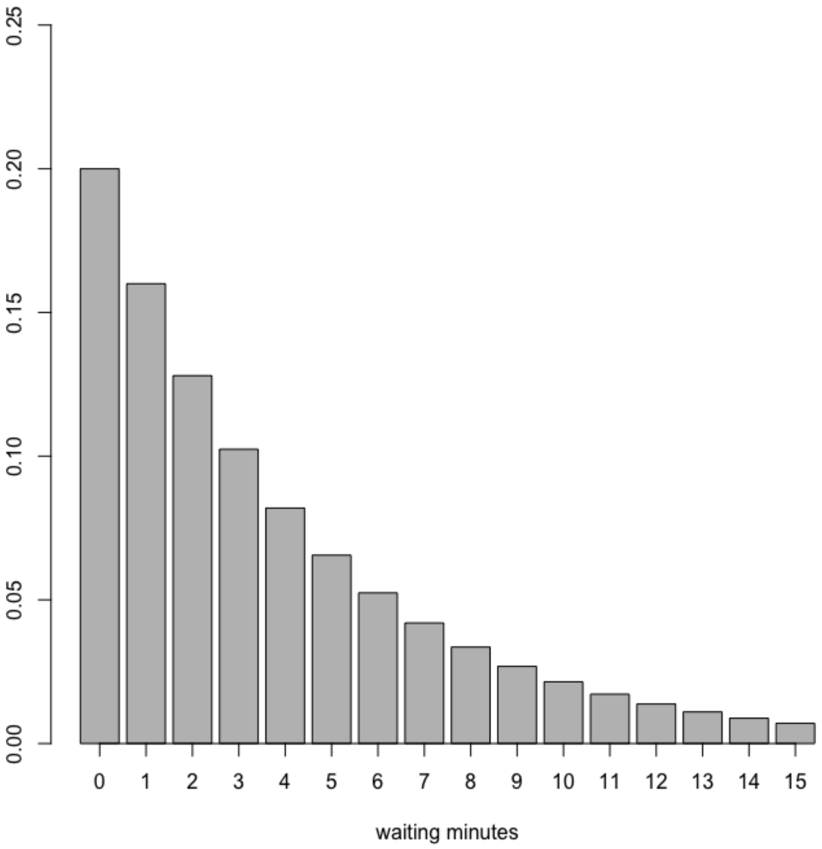
$$4 \exp(1 + \lambda_2) = \sum_{j=0}^{\infty} j \exp(-\lambda_1 j) = \frac{\exp(-\lambda_1)}{(1 - \exp(-\lambda_1))^2}.$$

Substituting the first of these into the latter, and solving for λ_1, we see that $\lambda_1 = \ln(5/4)$. Substituting that result back into the first equation, we further obtain $\exp(-1 - \lambda_2) = \frac{1}{5}$, so that

$$p_j = \exp(-1 - \lambda_2)\exp(-\lambda_1 j) = \frac{1}{5}\left(\frac{4}{5}\right)^j, j = 0, \ldots$$

It is easy to see that this defines a distribution; a "verification" is provided by the following code.

```
pmf_maxent <- function(x,lambda=4/5) (1-lambda)*(lambda)^x
sum(pmf_maxent(0:100))  # check if it's a distribution
mp <- barplot(pmf_maxent(0:15), ylim=c(0,.25),
              xlab="waiting minutes")
axis(1,at=mp,labels=paste(0:15))
```



This distribution could be used as a prior in a Bayesian analysis of the situation. Notice that some information about the data (in this case, only the sample mean) is used to define the MaxEnt prior.

Crucially, however, the data that is used to build the MaxEnt prior **cannot** be re-used as part of the likelihood computations. The situation is not unlike that of the training/testing paradigm of machine learning.

26.4 Posterior Distributions

The posterior distribution is used to estimate a variety of **model parameters of interest**, such as the mean, the median, the mode, etc.

It is possible to construct **credible intervals/regions** directly from the posterior (in contrast to the "confidence" intervals of frequentist inference). Given a posterior distribution on a parameter $\theta \in \mathbb{R}$, a $1-\alpha$ **credible interval** C.I. $[L, U]$ is an interval such that

$$P(L \leq \theta \leq U \mid D; I) \geq 1 - \alpha.$$

A similarly construction can be used for a joint credible region $\boldsymbol{\theta} \in \mathbb{R}^n$.

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.

26.4.1 High-Density Intervals

We can build the credible interval of θ-values using the **highest density interval** (HDI), i.e., We define a region C_k in parameter space with

$$C_k = \{\theta : P(\theta \mid D; I) \geq k\},$$

where k is the largest number such that

$$\int_{C_k} P(\theta \mid D; I)\, d\theta = 1 - \alpha.$$

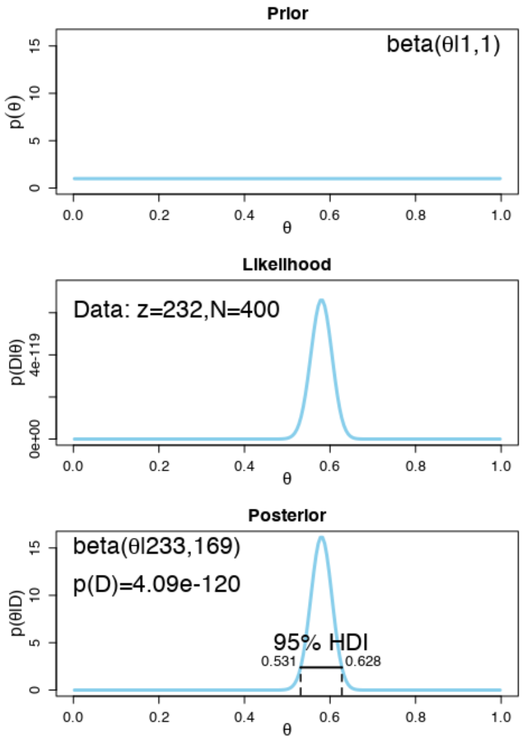
This typically has the effect of finding the **smallest region** C_k (in measure) region meeting the criterion.[21]

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

1. Suppose that we had no particular belief about the preference before the poll was published.[22] What is the 95% HDI on this belief after learning of the poll result?

   Solution: let preference for candidate A be denoted by 1, and preference for candidate B by 0. We can think of each voter's preference as arising from an independent Bernoulli trial.[23]

```
post = BernBeta(c(1,1), c(rep(1,232), rep(0,168)))
```



We see that the posterior distribution's 95% HDI ranges from 0.531 to 0.628, in favour of candidate A.

[21] The value k can be thought of the height of a horizontal line (or hyperplane, in the case of multivariate posteriors) overlaid on the posterior, whose intersection(s) with the latter define a region over which the integral of the posterior is $1 - \alpha$. In most cases, it must be found numerically.

[22] A non-informative uniform prior on the preference, which is to say, a Beta distribution with both parameters equal to 1.

[23] Assuming that the polled voters are selected randomly.

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

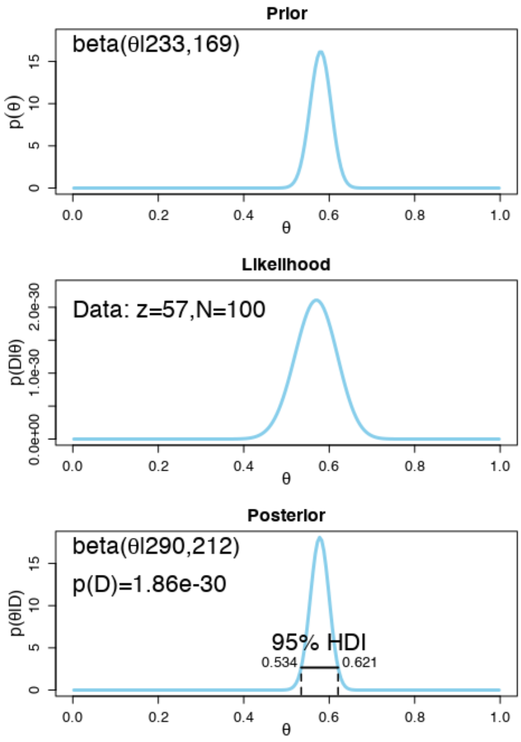
   Solution: the 95% HDI from the previous part shows that $\theta = 0.5$ is not among the credible values, hence it is not credible to believe that the population is equally divided in its preferences (at the 95%) level.

3. Say we conduct a follow-up poll to narrow our estimate of the population's preference. We randomly sample 100 people and find that 57 prefer candidate A. Assuming that the opinion of people has not changed between polls, what is the 95% HDI on the posterior?

   Solution: using the previous posterior as a new prior, we obtain the following results.

```
post = BernBeta( post, c(rep(1,57), rep(0,43)))
```

```
[1] 290 212
```



   The 95% HDI for the preference still leans towards candidate A, but is a bit narrower, ranging from 0.534 to 0.621.

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

   Solution: the 95% HDI from the previous results excludes $\theta = 0.5$; both the follow-up poll and the original poll suggest that the population is not equally divided (and actually prefers candidate A).

26.4.2 MCMC Methods

The true power of Bayesian inference is most keenly felt when the model specifications lead to a posteriors that cannot be manipulated analytically; in that case, it is usually possible to recreate a synthetic (or **simulated**) set of values that share the properties with a given posterior. Such processes are known as **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. The state of the chain after a number of steps is then used as a sample of the desired distribution.[24] MCMC techniques are often applied to solve integration and optimization problems in large-dimensional spaces.

These two types of problem play a fundamental role in machine learning, physics, statistics, econometrics and decision analysis. For instance, given variables $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and data D, the following (typically intractable) integration problems are central to Bayesian inference:

- **normalization** – in order to obtain the posterior $P(\boldsymbol{\theta} \mid D; I)$ given the prior $P(\boldsymbol{\theta} \mid I)$ and likelihood $P(D \mid \boldsymbol{\theta}; I)$, the normalizing (denominator) factor in Bayes' theorem needs to be computed

$$P(\boldsymbol{\theta} \mid D; I) = \frac{P(\boldsymbol{\theta} \mid I)P(D \mid \boldsymbol{\theta}; I)}{\int_{\boldsymbol{\Theta}} P(D \mid \boldsymbol{\theta}; I)P(\boldsymbol{\theta} \mid I)d\boldsymbol{\theta}};$$

- **marginalization** – given the joint posterior of $(\boldsymbol{\theta}, x)$, we may often be interested in the marginal posterior

$$P(\boldsymbol{\theta} \mid D; I) = \int P(\boldsymbol{\theta}, x \mid D; I)dx;$$

- **expectation** – the final objective of the analysis is often to obtain summary statistics of the form

$$E(f(\boldsymbol{\theta})) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{\theta})P(\boldsymbol{\theta} \mid D; I)d\boldsymbol{\theta}$$

  for some function of interest (i.e., $f(\boldsymbol{\theta}) = \boldsymbol{\theta}$ or $f(\boldsymbol{\theta}) = (\boldsymbol{\theta} - E(\boldsymbol{\theta}))^2$, which represent the mean and the variance, respectively).

26.4.3 The MH Algorithm

The Metropolis-Hastings (MH) algorithm is a specific type of Monte Carlo process; it is likely among the ten algorithms that have had the greatest influence on the development and practice of science and engineering in recent years.[25]

MH generates a random walk (that is, it generates a succession of posterior samples) in such a way that each step in the walk is **completely**

[24] The quality of the sample improves as a function of the number of steps.

[25] The celebrated **Gibbs sampler** can be viewed as a special case of MH.

independent of the preceding steps; the decision to reject or accept the proposed step is also independent of the walk's history.

Any process for which the current step is independent (forgetful) of the previous states, namely

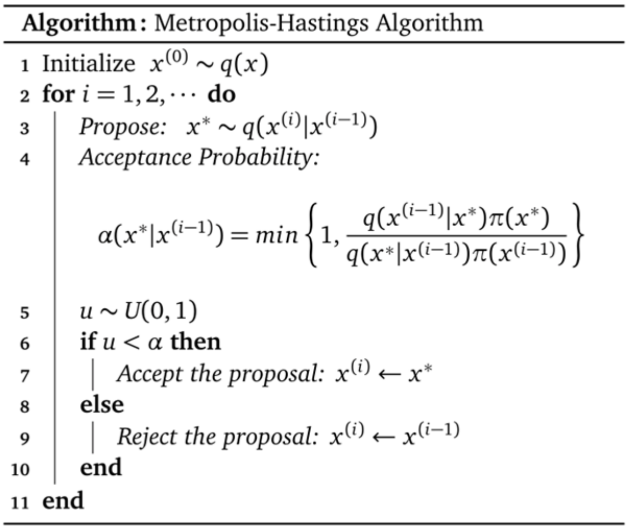
$$P(X_{n+1} = x \mid X_1 = x_1, \dots, X_n = x_n; I) = P(X_{n+1} = x \mid X_n = x_n; I)$$

for all n, X_j and x_j, $j = 1, \dots, n$, is called a **(first order) Markov process**, and a succession of such steps is a **(first order) Markov chain**.

MH uses a candidate or proposal distribution for the posterior, say $q(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters that is fixed by the user-called tuning parameters; MH then constructs a Markov Chain by proposing a value for $\boldsymbol{\theta}$ from this candidate distribution, and then either accepting or rejecting this value (with a certain probability).

Theoretically the proposal distributions can be nearly any distribution, but in practice it is recommended to keep to simple ones: 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), say.

The MH algorithm **simulates** samples from a probability distribution by making use of the full joint density function and (independent) proposal distributions for each of the variables of interest.



Algorithm: 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
```

The first step is to **initialize the sample value** for each random variable (often obtained by sampling from the variable's prior distribution). The main loop of the algorithm consists of three components:

1. **generate a candidate sample** x^* from the proposal distribution $q(x^{(i)} | x^{(i-1)})$;
2. **compute the acceptance probability** *via* the acceptance function $\alpha(x^* | x^{(i-1)})$ based on the proposal distribution and the full joint density $\pi(\cdot)$;
3. **accept the candidate sample** with probability α, the acceptance probability, or **reject it** otherwise.

Example (modified from [326, 328]): we use the MH algorithm to “learn” linear model parameters from a dataset. The **test data** for this example is generated as follows.

First, we establish the true model parameters.

```
set.seed(0)  # for replicability
t.A <- 10    # true slope
t.B <- 0     # true intercept
t.sd <- 20   # true noise
s.Size <- 50 # sample size
```

We will use equally spaced x values:

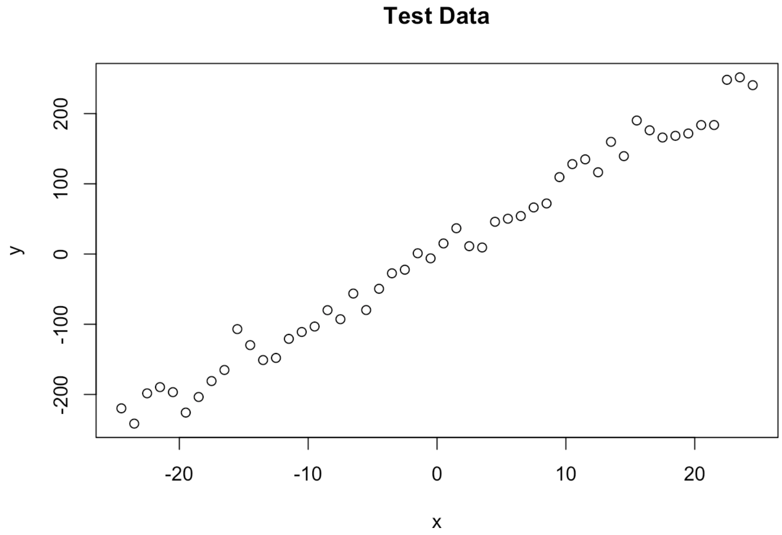
```
x <- (-(s.Size-1)/2):((s.Size-1)/2)
```

The corresponding y values are such that $y \sim \mathcal{N}(ax+b, \sigma^2)$:

```
y <-  t.A * x + t.B + rnorm(n=s.Size,mean=0,sd=t.sd)
```

The x values are balanced around zero in order to “de-correlate” the slope and the intercept.

```
plot(x,y, main="Test Data")
```



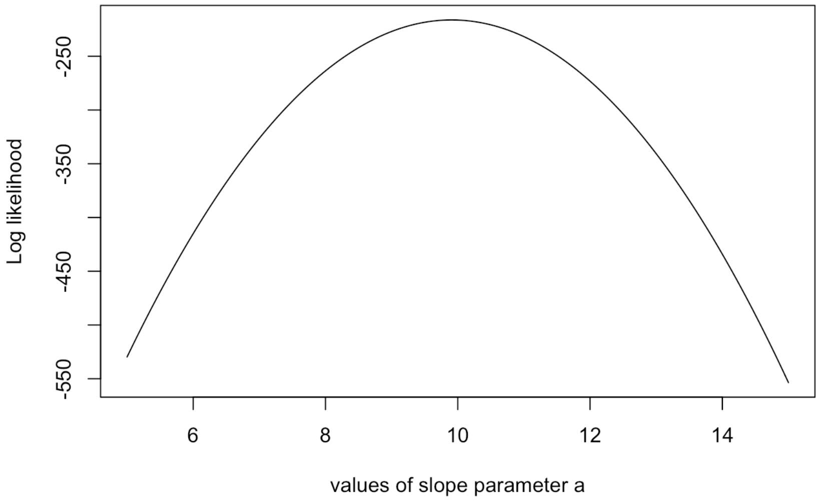
Defining the statistical model. The next step is to specify the statistical model. We already know that the data was created with a linear relationship $y = ax + b$ together with a normal error model $\mathcal{N}(0, \sigma^2)$, so we might as well use the same model for the fit and see if we can retrieve our original parameter values. Note however that, in general, the generating model is unknown.

Deriving the likelihood function from the model. A linear model of the form $y = ax + b + \mathcal{N}(0, \sigma^2)$ takes the parameters (a, b, σ) as **inputs**. The **output** should be the probability of obtaining the test data under this model: in this case, we only need to calculate the difference between the predictions $y = ax + b$ and the observed y, and then look up the probability (using dnorm) for such deviations to occur.

```
likehd <- function(param){
  a = param[1]
  b = param[2]
  sd = param[3]
  pred = a*x + b
  singlelikelihoods = dnorm(y, mean=pred, sd=sd, log=T)
  sumll = sum(singlelikelihoods)
  return(sumll)
}
```

For instance, we can find and plot the likelihood profile of the slope:

```
s.values <- function(x){return(likehd(c(x, t.B, t.sd)))}
s.likehds <- lapply(seq(1/2*t.A, 3/2*t.A, by=.05), s.values )
plot (seq(1/2*t.A, 3/2*t.A, by=.05), s.likehds , type="l",
  xlab = "values of slope parameter a", ylab = "Log likelihood")
```



Defining the priors. In Bayesian analysis, the next step is always required: we have to specify a prior distribution for each of the model parameters. To keep things simple, we will use a uniform distribution for the slope, and normal distributions for the noise and the intercept.[26]

```
prior <- function(param){
  a = param[1]
  b = param[2]
  sd = param[3]
  aprior = dunif(a, min=0, max=2*t.A, log = T)
  bprior = dnorm(b, mean=t.B, sd = 5, log = T)
  sdprior = dunif(sd, min=0, max=2*t.sd, log = T)
  return(aprior+bprior+sdprior)
}
```

[26] We will work with the logarithms of all quantities, so that the likelihood is a sum and not a product as would usually be the case.

The posterior. The product of prior by likelihood is the actual quantity that MCMC works with (it is not, strictly speaking, the posterior as it is not normalized).

```
posterior <- function(param){
  return (likehd(param) + prior(param))
}
```

Applying the MH algorithm. One of the most frequent applications of MH (as in this example) is sampling from the posterior density in Bayesian statistics.[27]

The aim of the algorithm is to jump around in parameter space, but in such a way as to have the probability to land at a point be proportional to the function we sample from (this is usually called the **target function**). In this case, the target function is the posterior that was defined previously.

This is achieved by

1. starting with a random parameter vector;
2. choosing a new parameter vector near the old value based on some probability density (the proposal function), and
3. jumping to this new point with a probability

   $$\alpha = \min\{1, g(\text{new})/g(\text{old})\},$$

   where g is the target.

The distribution of the parameter vectors MH visits converges to the target distribution g.

```
proposalfunction <- function(param){
  return(rnorm(3,mean = param, sd= c(0.1,0.5,0.3)))
}

run_metropolis_MCMC <- function(startvalue, iterations){
  chain = array(dim = c(iterations+1,3))
  chain[1,] = startvalue
  for (i in 1:iterations){
    proposal = proposalfunction(chain[i,])

    probab = exp(posterior(proposal) - posterior(chain[i,]))
    if (runif(1) < probab){
      chain[i+1,] = proposal
    }
    else{
      chain[i+1,] = chain[i,]
    }
  }
  return(chain)
}

startvalue = c(4,1,10) # random choice
chain = run_metropolis_MCMC(startvalue, 10000)
```

[27] The algorithm may be used to sample from any integrable function.

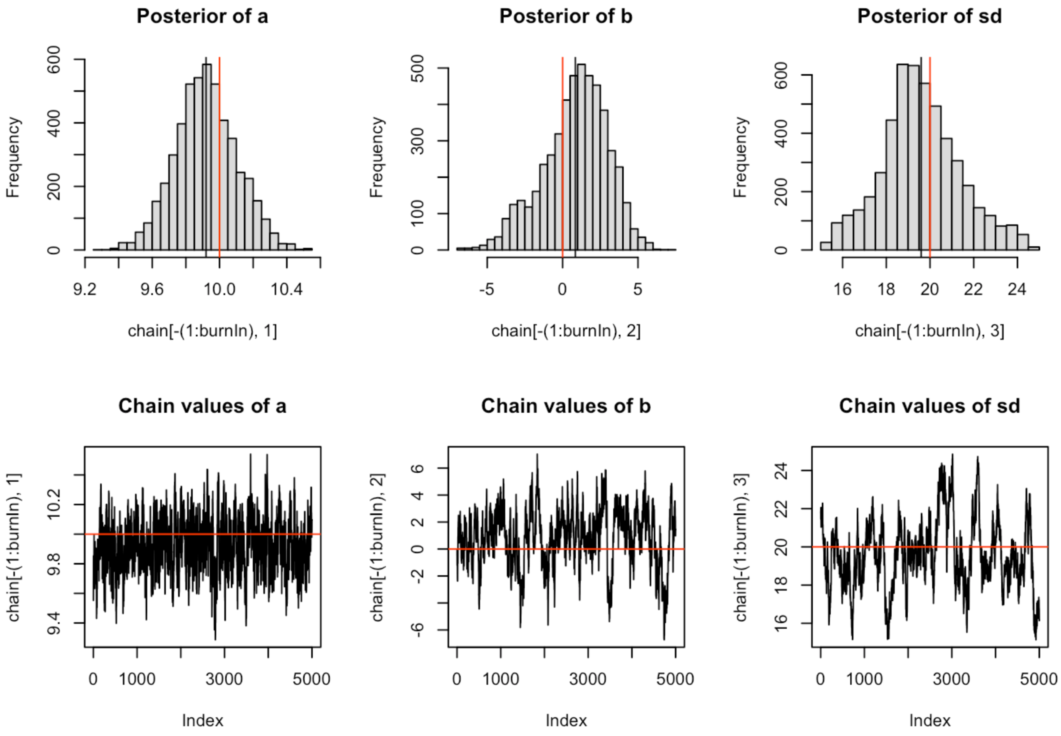
The first steps of the algorithm may be biased by the initialization process; they are usually discarded for the analysis (this is referred to as the **burn-in time**).

```
burnIn = 5000
acceptance = 1-mean(duplicated(chain[-(1:burnIn),]))
```

The **acceptance rate** is an interesting output to study: how often was a proposal rejected by the MH acceptance criterion? The acceptance rate can be influenced by the proposal function: generally, the nearer the proposal is to the latest value, the larger the acceptance rate.[28]

We plot the results below (the true parameter values are shown in red).

```
par(mfrow = c(2,3))
hist(chain[-(1:burnIn),1],nclass=30, main="Posterior of a")
abline(v = mean(chain[-(1:burnIn),1]))
abline(v = t.A, col="red" )
hist(chain[-(1:burnIn),2],nclass=30, main="Posterior of b")
abline(v = mean(chain[-(1:burnIn),2]))
abline(v = t.B, col="red" )
hist(chain[-(1:burnIn),3],nclass=30, main="Posterior of sd")
abline(v = mean(chain[-(1:burnIn),3]) )
abline(v = t.sd, col="red" )
plot(chain[-(1:burnIn),1], type = "l", main = "Chain values of a")
abline(h = t.A, col="red" )
plot(chain[-(1:burnIn),2], type = "l", main = "Chain values of b")
abline(h = t.B, col="red" )
plot(chain[-(1:burnIn),3], type = "l", main = "Chain values of sd")
abline(h = t.sd, col="red" )
```

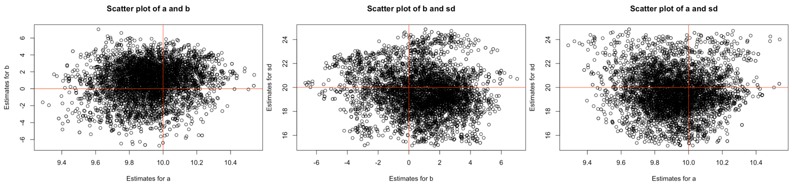


[28] Very high acceptance rates, however, are usually not beneficial, as this implies that the algorithms is "staying" in the same neighbourhood, which results in **sub-optimal probing of the parameter space** (there is very little **mixing**). Acceptance rates between 20% and 30% are considered optimal for typical applications [329].

The upper row shows posterior estimates for the slope a, intercept b, and standard deviation of the error σ; the lower row shows the Markov Chain of parameter values. We retrieve (more or less) the original parameters that were used to create the data, and there is a certain area around the highest posterior values that also show some support by the data, which is the Bayesian equivalent of confidence intervals.

These posterior distributions are **marginal distributions**; the pairwise joint distributions are shown below (again, with true parameter values in red – the horizontal and vertical lines).

```
plot(chain[-(1:burnIn),1:2], main="Scatter plot of a and b",
     xlab="Estimates for a", ylab="Estimates for b")
abline(v = t.A, col="red" )
abline(h = t.B, col="red" )
plot(chain[-(1:burnIn),2:3], main="Scatter plot of b and sd",
     xlab="Estimates for b", ylab="Estimates for sd")
abline(v = t.B, col="red" )
abline(h = t.sd, col="red" )
plot(chain[-(1:burnIn),c(1,3)], main="Scatter plot of a and sd",
     xlab="Estimates for a", ylab="Estimates for sd")
abline(v = t.A, col="red" )
abline(h = t.sd, col="red" )
```



The posterior distributions certainly do seem to contain the true parameter values. By way of comparison, a simple linear regression analysis would yield the following estimates:

```
summary(lm(y~x))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-33.067 -12.201  -3.733  14.562  46.192

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4786     2.6115   0.183    0.855
x             9.9082     0.1810  54.751   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 18.47 on 48 degrees of freedom
Multiple R-squared:  0.9842,    Adjusted R-squared:  0.9839
F-statistic:  2998 on 1 and 48 DF,  p-value: < 2.2e-16
```

Which method is best, in this context? What are the advantages and disadvantages of each?

26.5 Additional Topics

According to [330],

> the central feature of Bayesian inference is the direct quantification of uncertainty.

Bayesian approach to modeling uncertainty is particularly useful when:

- the available data is **limited**;
- there is some concern about **overfitting**;
- some facts are **more likely to be true** than others, but that information is not contained in the data, or
- the precise likelihood of certain facts is more important than solely determining which fact is most likely (or least likely).

As discussed previously, Bayesian methods have a number of powerful features. They allow analysts to:

- incorporate **specific knowledge** about parameters of interest;
- logically **update knowledge** about the parameter after observing sample data;
- make **formal probability statements** about parameters of interest;
- specify **model assumptions** and check model quality and sensitivity to these assumptions in a straightforward manner, and
- provide **probability distributions** rather than point estimates.

26.5.1 Uncertainty

The following example represents a Bayesian approach to dealing with the uncertainty of the so-called **envelope paradox**.

Example: you are given two indistinguishable envelopes, each containing a cheque, one being twice as much as the other. You may pick one envelope and keep the money it contains. Having chosen an envelope at will, but before inspecting it, you are given the chance to switch envelopes. Should you switch? What is the expected outcome in doing so? Explain how this game leads to infinite cycling.

Solution: let V be the (unknown) value found in the envelope after the first selection. The other envelope then contains either $\frac{1}{2}V$ or $2V$, both with probability 0.5, and the expected value of trading is

$$E[\text{trade}] = 0.5 \times \frac{1}{2}V + 0.5 \times 2V = \frac{5}{4}V > V;$$

and so it appears that trading is advantageous.

Let the (still unknown) value of the cheque in the new envelope be W. The same argument shows that the expected value of trading *that* envelope is $\frac{5}{4}W > W$, so it would make sense to trade the envelope once more, and yet once more, and so on, leading to infinite cycling.

There is a Bayesian approach to the problem, however. Let V be the (uncertain) value in the original selection, and W be the (also uncertain) value in the second envelope. A proper resolution requires a joint (prior) distribution for V and W. Now, in the absence of any other information, the most we can say about this distribution using the maximum entropy principle is that $P(V < W) = P(V > W) = 0.5$.

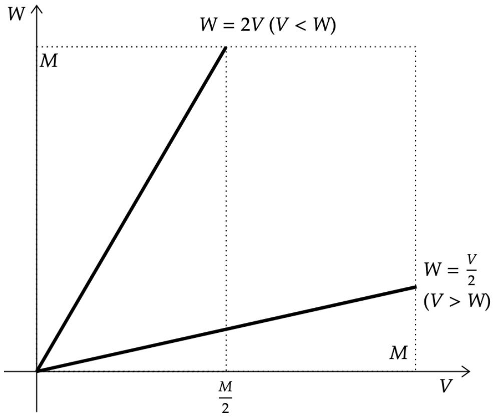
By definition, if $V < W$, then $W = 2V$; if, on the other hand, $V > W$ then $W = \frac{V}{2}$. We now show that the expected value in both envelopes is the same, and thus that trading envelope is no better strategy than keeping the original selection. Using Bayes' Theorem, we compute that

$$\begin{aligned} E[W] &= E[W|V < W]P(V < W) + E[W|V > W]P(V > W) \\ &= E[2V|V < W] \cdot 0.5 + E[0.5V|V > W] \cdot 0.5 \\ &= E[V|V < W] + 0.25 \cdot E[V|V > W], \end{aligned}$$

while

$$\begin{aligned} E[V] &= E[V|V < W]P(V < W) + E[V|V > W]P(V > W) \\ &= 0.5 \cdot E[V|V < W] + 0.5 \cdot E[V|V > W]. \end{aligned}$$

Before we can proceed any further, we must have some information about the joint distribution $P(V, W)$ (note, however, that $E[W]$ will not typically be equal to $\frac{5}{4}V$, as had been assumed at the start of the solution). The domain Ω of the joint probability consists of those pairs (V, W) satisfying $V = 2W$ $(V > W)$ or $W = 2V$ $(V < W)$ for $0 < V, W < M$, where $M < \infty$ is some upper limit on the value of each cheque.[29]



We have assumed that the probability weight on each branch of Ω is $1/2$; if we further assume, say, that the cheque value is as likely to be any of the allowable values on these branches, then the joint distribution is

$$P(V, W) = \begin{cases} \frac{1}{M} & \text{if } V < W \\ \frac{1}{2M} & \text{if } V > W \\ 0 & \text{otherwise} \end{cases}$$

and the expectations listed above are

$$E[V|V < W] = \int_{V<W} V \cdot P(V, W)\, d\Omega = \int_0^{M/2} V \cdot \frac{1}{M}\, dV = \frac{M}{8}$$

and

$$E[V|V > W] = \int_{V>W} V \cdot P(V, W)\, d\Omega = \int_0^{M} V \cdot \frac{1}{2M}\, dV = \frac{M}{4}.$$

[29] In the worst case scenario, M would have to be smaller than the total amount of wealth available to humanity throughout history, although in practice M should be substantially smaller. Obviously, a different argument will need to be made in the case $M = \infty$.

Therefore,

$$E[W] = \frac{M}{8} + 0.25 \cdot \frac{M}{4} = \frac{3M}{16}$$

and

$$E[V] = 0.5 \cdot \frac{M}{8} + 0.5 \cdot \frac{M}{4} = \frac{3M}{16},$$

and switching the envelope does not change the expected value of the outcome. There is no paradox; no infinite cycling.

Example: After the sudden death of her two baby sons, Sally Clark was sentenced by a U.K. court to life in prison in 1996. Among other errors, expert witness Sir Roy Meadow had wrongly interpreted the small probability of two cot deaths as a small probability of Clark's innocence. After a long campaign, which included the refutation of Meadow's statistics using Bayesian statistics, Clark was released in 2003. While Clark's innocence could not be proven beyond the shadow of a doubt using such methods, her culpability could also not be established beyond reasonable doubt and she was cleared.[30]

26.5.2 Bayesian A/B Testing

A/B **testing** is an excellent tool for deciding whether or not to roll out incremental features. To perform an A/B test, we divide users randomly into a **test** group and into a **control** group, then provide the new feature to the test group while letting the control group continue to experience the current version of the product.

If the randomization procedure is appropriate, we may be able to attribute any difference in outcomes between the two groups to the changes we are rolling out without having to account for other sources of variation affecting the user behaviour. Before acting on these results, however, it is important to understand the likelihood that any observed differences is merely due to chance rather than to product modification.

For example, it is perfectly possible to obtain different H/T ratios between two fair coins if we only conduct a limited number of tosses; In the same manner, it is possible to observe a change between the A and B groups even if the underlying user behavior is identical.

Example: (modified from [332]) *Wakefield Tiles* is a company that sells floor tiles by mail order. They are trying to become an active player into the lucrative Chelsea market by offering a new type of tile to the region's contractors.

The marketing department have conducted a pilot study and tried two different marketing methods:

- A – sending a colourful brochure in the mail to invite contractors to visit the company's showroom;
- B – sending a colourful brochure in the mail to invite contractors to visit the company's showroom, while including free tile samples.

The marketing department sent out 16 mail packages of type A and 16 mail packages of type B. Four Chelseaites that received a package of type A visited the showroom, while 8 of those receiving a package of type B did the same.

[30] An informative write-up of the situation can be found online [331].

The company is aware that:

- a mailing of type A costs 30$ (printing cost and postage);
- a mailing of type B costs 300$ (also includes the cost of the free tile samples);
- a visit to the showroom yields, on average, 1000$ in revenue during the next year.

Which of the methods (A or B) is most advantageous to *Wakefield Tiles*?

Solution: the Bayesian solution requires the construction of a prior distribution and of a **generative model**; as part of the generative model, we will need to produce n replicates of samples from the binomial distribution.[31]

The binomial distribution simulates `n` times the number of "successes" when performing `size` trials (mailings), where the probability of a "success" is `prob`. A commonly used prior for `prob` is the uniform distribution $U(0, 1)$, from which we sample in `R` *via* `runif(1, min = 0, max = 1)`.

We start by setting a seed for replicability, and set the number of replicates (trials).

```
set.seed(1111) # for replicability
n.draws <- 200000
```

Next, we generate a probability of success for mailings A and B, for each of the replicates.

```
prior <- data.frame(p.A = runif(n.draws, 0, 1),
                    p.B = runif(n.draws, 0, 1))
```

The generative model tells us how many visitors to expect for mailing types A, B, for each replicate.

```
generative.model <- function(p.A, p.B) {
  visitors.A <- rbinom(1, 16, p.A)
  visitors.B <- rbinom(1, 16, p.B)
  c(visitors.A = visitors.A, visitors.B = visitors.B)
}
```

We then simulate data using the parameters from the prior and the generative model. This yields the actual number of visitors for each replicate.

```
sim.data <- as.data.frame( t(sapply(1:n.draws, function(i) {
  generative.model(prior$p.A[i], prior$p.B[i])})))
```

Only those prior probabilities for which the generative model match the observed data are retained.

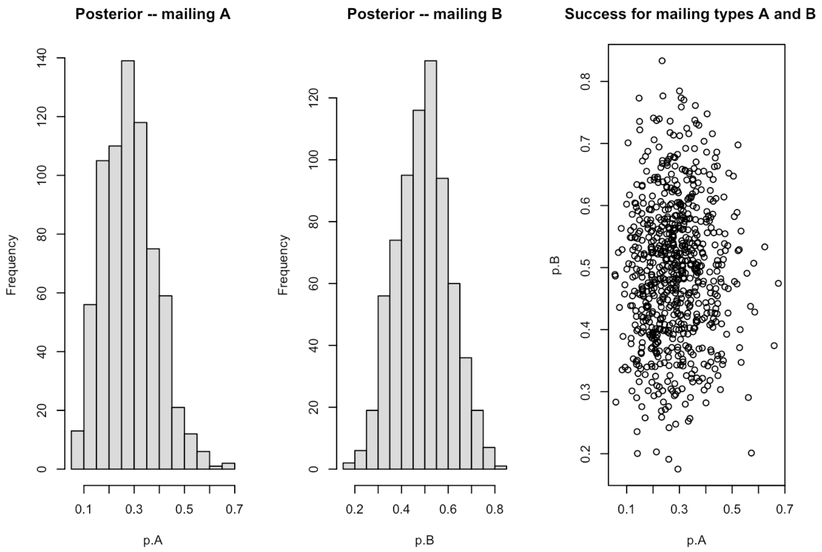
[31] Which can be achieved in `R` using `rbinom(n,size,prob)`.

```
posterior <- prior[sim.data$visitors.A == 4 &
                   sim.data$visitors.B == 8, ]
```

In this case, there are enough trials to ensure that the posterior is non-empty; what could be done if that was not the case?

Finally, we visualize the posteriors:

```
par(mfrow = c(1,3))
hist(posterior$p.A, main = "Posterior -- mailing A",
     xlab="p.A")
hist(posterior$p.B, main = "Posterior -- mailing B",
     xlab="p.B")
plot(posterior,main = "Success for mailing types A and B",
     xlab="p.A", ylab="p.B")
```

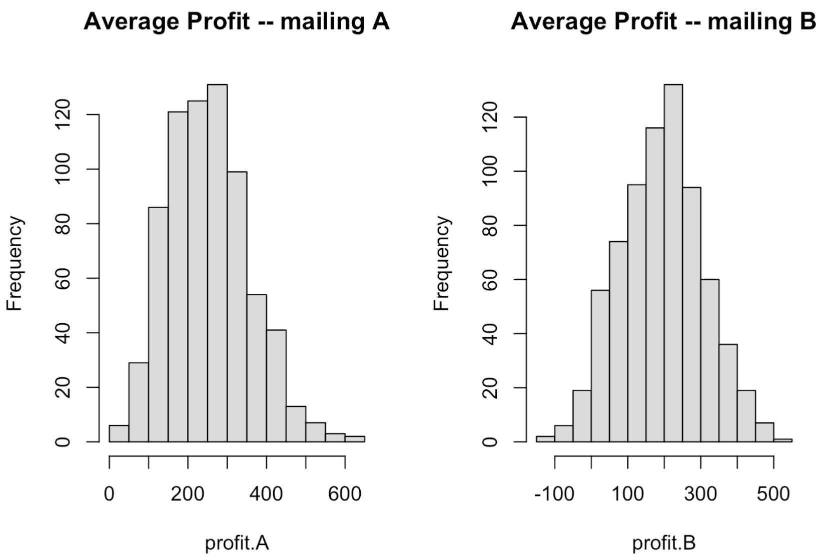


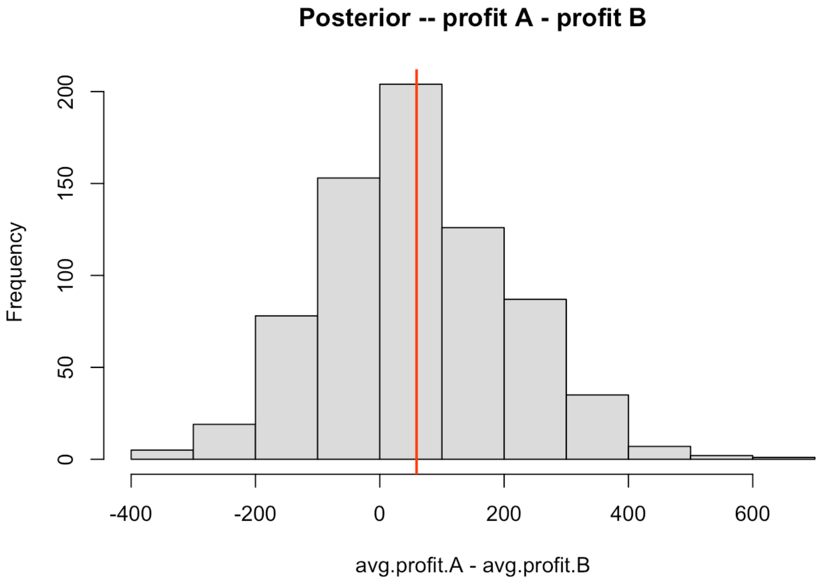
The posterior distributions for the probability of success for each mailing types are produced as below (see next page for display).

```
par(mfrow = c(1,2))
avg.profit.A <- -30 + posterior$p.A * 1000
avg.profit.B <- -300 + posterior$p.B * 1000
hist(avg.profit.A, main = "Average Profit -- mailing A",
     xlab="profit.A")
hist(avg.profit.B, main = "Average Profit -- mailing B",
     xlab="profit.B")
```

In order to estimate the average profit for each mailing type, we use the posterior distributions for the probability of success (see next page).

```
hist(avg.profit.A - avg.profit.B, main="Posterior --
     profit A - profit B")
(expected.avg.profit.diff <- mean(avg.profit.A - avg.profit.B))
abline(v = expected.avg.profit.diff , col = "red", lwd =2)
```





```
[1] 59.13869
```

The expected profit for mailing type A is about 60$ higher than for mailing type B (numbers may vary, depending on the seed). Keeping it simple seems to be a better idea in this context.

26.6 Exercises

1. Cognitive neuroscientists investigate which areas of the brain are active during particular mental tasks. In many situations, researchers observe that a certain region of the brain is active and infer that a particular cognitive function is therefore being carried out; [333] cautioned that such inferences are not necessarily firm and need to be made with Bayes' rule in mind. The same paper reports the following frequency table of previous studies that involved any language-related task (specifically phonological

and semantic processing) and whether or not a particular **region of interest** (ROI) in the brain was activated (see table below). Suppose that a new study is conducted and finds that the ROI is activated (A). If the prior probability that the task involves language processing is $P(L) = 0.5$, what is the posterior probability, $P(L \mid A)$, given that the ROI is activated?

	Language (L)	Other ($\overline{L}$)
Activated (A)	166	199
Not Activated ($\overline{A}$)	703	2154

2. Suppose that, in 1975, 52% of UK voters supported the Labour Party and 48% the Conservative Party. Suppose further that 55% of Labour voters wanted the UK to remain part of the EEC and 85% of Conservative voters were also in favour. What is the probability that a person voting "Yes" (in favour of remaining in the EEC) in the 1975 referendum is a Labour voter? [334]
3. Given the following statistics, what is the probability that a woman over 50 years of age has breast cancer if she receives a positive mammogram result? [Bayes' Theorem Problems, Definition and Examples]
   - 1% of women over 50 have breast cancer;
   - 90% of women over 50 who have breast cancer test positive on mammograms.
   - 8% of women over 50 will obtain a false positive result on a breast cancer test.
4. 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? (Go Sens!)
   - your belief in the effectiveness of homeopathic remedies?
   - your belief in the effectiveness of Bayesian analysis?
5. Suppose that a test for a particular disease has a very high success rate. When a patient has the disease, the test accurately reports a 'positive' with probability 0.99; when they do not, 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? Is this problem any different from problem 3?
6. A road safety analyst has access to a dataset of fatal vehicle collisions (such as Canada's *National Collision Database*) 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. 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. The analyst will consider 3 propositions:

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

The analysts may have an interest in $P(B \mid A; I), P(C \mid A; I), P(B, C \mid A; I)$, $P(B, -C \mid A; I)$, or $P(-B, C \mid A; I)$. Derive an expression to compute the probability that "bad" weather and road conditions were present at the time of the collision.

7. A **Mild Winter** scenario (we use the set-up of question 6): 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?
8. **Not Quite as Mild a Winter** scenario (we use the set-up of questions 6 and 7): 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 7?
9. Use the set-up of questions 6-8. 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?
10. Use the set-up of questions 6-9. 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.
11. **Really Rough Winter** scenario (see questions 6-10): 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: $k = 0.1, 1, 10, 100$. Which of these scenarios is most likely?
12. Use the set-up of questions 6-11. 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.

13. Use the set-up of questions 6-12. Repeat the process for the other conditional probabilities of interest.
14. A lifetime's supply of poutine is placed randomly behind one of three identical doors. The other two doors lead to empty rooms. 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?

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

15. How many heads in a row would you need to observe before you would start doubting whether a coin is fair or not?
16. Estimate the parameters (μ_i, σ_i) for $i = 1, \ldots, 12$ in the Salary example.
17. Play with the parameters and implement new scenarios for the Money (Dollar Bill Y'All) example.
18. Play with the `BernBeta()` function. Do you spot anything surprising?
19. Suppose you have in your possession a coin that you know was minted by the federal government and for which you have no reason to suspect tampering of any kind. Your prior belief about fairness of the coin is thus strong. You flip the coin 10 times and record 9 H(eads). What is your predicted probability of obtaining 1H on the 11th flip? Explain your answer carefully; justify your choice of prior. How would your answer change (if at all) if you use a frequentist viewpoint?
20. A mysterious stranger hands you a different coin, this one made of some strange-to-the-touch material, on which the words "Global Tricksters Association" You flip the coin 10 times and once again record 9H. What is your predicted probability of obtaining 1H on the 11th flip? Explain your answer carefully; justify your choice of prior. Hint: what would be a reasonable prior for this scenario?
21. A group of adults are doing a simple learning experiment: when they see the two words "radio" and "ocean" appear simultaneously on a computer screen, they are asked to press the F key on the keyboard; whenever the words "radio" and "mountain" appear on the screen, they are asked to press the J key. After several practice repetitions, two new tasks are introduced: in the first, the word "radio" appears by itself and the participants are asked to provide the best response (F or J) based on what they learned before; in the second, the words "ocean" and "mountain" appear simultaneously and the participants are once again asked to provide the best response. This is repeated with 50 people. The data shows that, for the first test, 40 participants answered with F and 10 with J; while for the second test, 15 responded with F and 35 with J. Are people biased toward F or toward J for either of the two tests? To answer this question, assume a uniform prior, and use a 95% HDI to decide which biases can be declared to be credible.
22. Suppose that the marketing group of a company is testing a new web page, with the hope of increasing the conversion rate (proportion of visitors who sign up or take some other action). The data is

collected in the file `ab_data.csv`, which lists user visits with whether they were sent to the new page or the old page, and whether there was a conversion.

   a) Explore and visualize the dataset.
   b) We conduct Bayesian A/B testing, by defining and updating independent priors on the old and new conversion rates, to arrive at respective posterior distributions for the old page and the new page. Try a prior of `Beta(alpha=2, beta=20)` for the old rate, which represents what has been observed in the past. Start with a subset of 100 data points and perform inference. Find the posterior probability that the new page has a higher conversion rate. Hint: use random samples from the independent posteriors to estimate the probability. Update the posteriors with another 100 data points. At what data size do the priors become irrelevant?

23. Sometimes we don't just want to estimate a dependent variable, we want a probability distribution for it. For instance, if one's life expectancy is 80 years, we might want to know whether it's a 50/50 spli between 0 years and 160 years, or some other distribution.
    a) Load the `mimic3d.csv` dataset which lists the length of stay in a hospital (`LOSdays`) along with a number of other variables. Explore and visualize this dataset.
    b) Construct a dataset `patients.csv` containing information about 10 or so (or more) "patients", for all but the `LOSdays` variable (you may use friends and family members, classmates, etc. as a basis for your observations).
    c) Predict the length of the hospital stay for the patients in the dataset by conducting a Bayesian linear regression analysis. What's the probability of staying longer than 2 days and therefore definitely missing work? Use normal priors for simplicity.