This document is an excerpt from
Resampling Stats in MATLAB
Daniel T. Kaplan
Copyright (c) 1999 by Daniel T. Kaplan, All Rights Reserved
This document differs from the published book in pagination and in the omission (unintentional, but unavoidable for technical reasons) of figures and cross-references from the book. It is provided as a courtesy to those who wish to examine the book, but not intended as a replacement for the published book, which is available from
Resampling Stats, Inc.
www.resample.com
703-522-2713

Chapter 4: Updating our View: Bayesian Analysis

The way we view the world is often stated in terms of probabilities: the probability of drawing an ace of spades from a normal deck of cards is $\frac{1}{52}$; the probability of getting lung cancer is 7%; the chance of rain today is 40% according to the television report last night, but 80% according to the morning radio news.

Actually, the above statements are not complete: they fail to tell us the complete situation to which they apply. If you are playing stud poker, the probability of drawing the ace of spades changes depending on how many cards have been dealt and on the information you already have: if you already hold the ace of spades your opponent has zero chance of getting that card, although your opponent may reckon the chances differently based on her own information or lack thereof. The probability of getting lung cancer depends critically on whether a person smokes. For those who do not smoke, the probability of getting lung cancer is less than 1%; for those who do smoke, the probability is much greater than 7%. The weather forecast is based on reading current weather conditions — the locations of cold and warm fronts, etc. If those conditions change, then the forecast probability changes.

Data tell us something about the world. One way to think about data is that they provide information that may change our description of the world in terms of probabilities: a change in the position of a cold front can change the forecast probability of rain.

Suppose we are in the process of collecting some new data, but have not yet examined these data. Our view of the world *before* we consider our new data is our *prior* view. Our view after we take our data into account is our *posterior* view. Data provide the bridge between the prior view and the posterior view. Resampling provides one way of computing posterior probabilities given our data and our prior probabilities.

Example 1: The Basketball Slump Revisited

Consider again Larry Bird's performance in the 1988 playoffs, which we looked at in Example **??** using the hypothesis testing framework. Here, we'll take a Bayesian perspective.

Before the playoffs start, we have some opinion about the chance that Bird will be in a slump during the playoffs. This opinion might be founded on our previous observations of Bird's playoff performances in other years, his playing record just before the playoff began, whether he has had an injury recently, whether he has just signed a lucrative contract, or whatever other factors we deem relevant. Whatever the origins of our opinion, informed or not, subjective or not, this constitutes our prior view of the probability that Bird will be in a slump during the playoffs.

Three games later, we have some new data. Bird has scored 20 out of 57 attempts. This performance might alter our view about whether Bird is in a slump or not — it would seem reasonable to change our opinion in the direction of increasing the probability that he is in a slump during the playoffs. But how can the appropriate change be calculated from the data?

In order to talk about change, we need to consider "change from what?" We start by describing the prior probability distribution. Rather than dividing the possibilities into slump or no-slump, we hypothesize a distribution for Bird's success rate in shooting baskets:

```
prior = [5 .25; 5 .30; 5 .35; 5 .40; 5 .45; 75 .52 ];
```

This says that there is a 5% chance that Bird is shooting quite poorly (for him) — making 25 out of 100 shots. Similarly, there is a 5% chance of him making 30 out of 100 shots and so on. This distribution results in Bird shooting, on average, about 48 out of 100 shots, which is consistent with his long-term record. The other details of the distribution are somewhat arbitrary, but they are our best guess about how Bird shoots.

How did we decide on the particular probabilities and values to use in `prior`? To be honest, we made them up. Perhaps they reflect Bird's shooting record in regular season games; perhaps they incorporate some other information. As we said above, they are our "best guess" about Bird's shooting patterns.

In the program `birdbayesian.m`, we use this prior probability distribution to generate many simulations of Bird's taking 57 shots. This is done using a two-stage process. First, we pick a shooting rate from the prior distribution:

```
rate = sample(1,prior);
```

Next, we generate shots according to the rate we have just picked. Given that `rate` is the probability of a successful basket, then `1-rate` is the probability of missing. We'll code a miss as `0` and a success as `1`, counting how many successes occur in 57 consecutive samples:

```
bird = [(1-rate) 0; rate 1];
baskets = count( 1 == sample(57, bird) );
```

This constitutes a single simulation of 57 consecutive shots at the given success rate. We can carry out many such simulations, tallying the result of each simulation. We'll keep track both of the success rate randomly selected from the prior distribution and the corresponding number of baskets in the simulated 57-shot sequence at that success rate.

```
tally rate raterecord;
tally baskets scorerecord;
```

So far, all we have are many simulations of the success rate of the 57-shot sequence. The simulations are consistent with our prior probability distribution.

Now, we want to incorporate the knowledge that comes from our observed data that Bird made 20 of the 57 shots. Only those simulations where Bird scored exactly 20 shots are consistent both with our prior distribution and with the observed data. Therefore, we select out those shooting rates that resulted in observing exactly 20 baskets:

```
≫  posterior = raterecord( find( scorerecord == 20 ));
```

This give a new set of data, `posterior`, that reflects the relative probabilities of different shooting rates that are consistent both with the prior probability distribution and the observed data: this is our posterior probability distribution.

Figure 1:
The Bayesian analysis of Bird's playoff performance. Top: our prior distribution of Bird's success rate. Bottom: The posterior distribution conditioned on the observation that Bird made 20 out of 57 shots.

If we define a slump (for Bird) as a success rate less than 40%, we can compute the posterior probability of being in a slump:

$\gg$ `proportion(posterior<.40)` $\Rightarrow$ *`ans: .51`*

Thus, we find that there is about a 50% probability that Bird was in a slump for the playoffs. This result reflects both the observed data and our subjective prior probability — before the data were collected — that Bird was in a slump.

If you are concerned about how the details of `prior` influenced the results of the calculation, try changing some of those details to see how they reflect the results. You can even try making several different priors and assign each of them a subjective probability of being correct. (See the documentation for URN for programming details.) Insofar as your results depend on a detail of your prior which cannot be justified, your results cannot be justified. But it often happens that the details of the prior have little effect on the results, in which case it doesn't matter whether the detail is justified or not.

Example 2: Revisiting the Space Shuttle

In designing the space shuttle, engineers did a careful job to estimate risks. Extensive studies were done to estimate the probabilities of failure of each component. For some systems such as the in-flight computers, the probability of component failure was substantial. Sufficient backup systems were provided so that the overall system had a low probability of failure.

In estimating the probability of an accident, certain assumptions are made. For instance, testing of the now-infamous O-ring seals was done at typical temperatures encountered at the launch site in Florida. On the day of the Challenger disaster, however, temperatures were well below normal. (See Tufte [?] for a case study of how data presentation techniques used by NASA engineers unintentionally obscured the relationship between temperature and O-ring failure.)

We saw in Example **??** that on the eve of the Challenger flight the 95% confidence interval for the accident rate as 0% to 11%. The upper end of this interval hardly seems to justify a decision to class the launch as non-experimental or safe. Remember, though, that there was more information available than the observed data of no crashes in 24 flights. There were also the engineering studies that showed a very low risk of accident.

The engineering studies can be used to estimate a prior probability distribution for the accident rate. For the sake of this example, we will assume this engineering-based prior to be an accident rate uniformly distributed in the range 0% to 0.01%. This is a very low accident rate as befits careful engineering design. But there is also the possibility that the assumptions that went into the engineering analysis have been violated.

It's impossible to give a completely objective number for the probability that the actual conditions critically violate engineering assumptions. But it seems prudent — on the first flight — to assume that this probability is quite high, say, 75%. If the engineering assumptions are invalid, we do not know what the accident rate is, so for want of other information we will assume that it is equally likely to be anything from 0 to 100%.

Overall, our prior distribution for the accident rate is

```
% engineering-based accident rate
engineering = uniform(0, 0.0001);
% accident rate from failed assumptions
failed = uniform(0,1);
 % combine these two rates, using our
% assumed 25% chance that the engineering
% assumptions are satisfied.
prior = urn(.25, engineering, .75, failed);
```

After the first flight we have some data about the accident rate. We can use these data to convert the prior distribution to a posterior distribution.

```
function posterior = updateshuttle(prior, crash)
% posterior = updateshuttle(prior, crash)
% translate a prior accident rate into a posterior
% prior -- prior probability distribution of accident rate
% crash -- 0 if there was no crash, 1 otherwise.
% The posterior takes the form of a list of accident rates.
```

```
Ntrials = 1000;
posterior = starttally;
for trial = 1:Ntrials
   rate = sample(1, prior);
   thisflight = sample(1, [(1-rate) 0; rate 1]);
   % 0 means no accident, 1 == accident
   if thisflight == crash % observation matches simulation
      posterior = [posterior rate]; %or: tally rate posterior;
   end
end
```

Figure 2:
Probability distributions for the accident rate for the Space Shuttle. The logarithm of the accident rate is shown. (When the logarithm is -1, for example, this means that the accident rate is 0.1.) The prior distribution is based on engineering information and a subjective assumption that — before any flights occur — the engineering assumptions are quite likely to be wrong. The posterior is computed after 1, 5, 10, and 20 flights, with no accidents observed.

To find the posterior after the first flight, where the observation is that there was not a crash:

```
≫  pos1 = updateshuttle(prior,0);
```

For the second flight, our prior is the posterior distribution from after the first flight.

```
≫  pos2 = updateshuttle(pos1,0);
```

Similarly, the prior before the third flight is the posterior for the second flight, and so on.

Figure **??** shows the distribution of the accident rate for the prior and for the posterior after 1 flight, 5 flights, 10 flights, and 20 flights, with no crashes observed. Since the engineering-based accident rates are so much smaller than the assumption-violation accident rates, the plots show the logarithm of the accident rate. For example, a bar in the plot near -5 on the x-axis corresponds to an accident rate of 10^{-5} or 0.001%.

In the histograms in Fig. **??** there are two broad peaks. The rightmost peak (at large accident rates) corresponds to the accidents that arise from violation of the engineering assumptions. The leftmost peak corresponds to accidents anticipated by the engineering calculations. The plots show that as the number of incident-free flights increases, the distribution of accident rates shifts strongly to the left, toward the engineering calculations. The experience with 20 flights provides evidence that lowers the probability of an assumption-violation accident from the initial prior of 75% to an after-20-flight posterior of 17%.

In reality, NASA has a lot of other information beyond whether or not there was a crash. This information comes from the telemetered systems on the shuttle. This information could be used to revise the engineering estimates and to reduce the probability of the possible unknown violations of the engineering assumptions. These calculations could be made using the same Bayesian approach taken in this example:

$$\begin{array}{ccc} & \text{data} & \\ & \downarrow & \\ \text{prior} & \rightarrow & \text{posterior} \end{array}$$