Statistical Rethinking for Soil, Water, and Fish

Josh Erickson

2023-11-23

Contents

Chapter 1

The Golem of Prague

Right now I'm just trying to work through what seems to be 4 years of reading statistical rethinking and learning statistics. Bare bones right now but hoping to make this more of a thing moving forward. Going to try and keep the context within a natural resource framework (fisheries, hydrology, soils).

I won't go into Chapter 1 because it's mostly for you to read on your own. Great content and outline for the rest of the book but not much to put on here.

Chapter 2

Small Worlds and Large Worlds

2.1 Small Worlds and Large Worlds

A great distinction between the small and large world is the *small world* is the model itself and the *large world* is the world we hope to deploy in. This is the challenge of statistical modeling and is elevated by forgetting this distinction.

In the small world there are no surprises and it is important to verify the logic, making sure that it performs as expected under favorable assumptions.

The large world has events that were not expected or imagined in the small world, e.g. coupling of events, a priori understanding is misleading, etc. This is essentially the modeling adage 'all models are wrong, but some are useful'. Just because the model in the small world makes logical sense doesn't mean that it will be consistent in the large world. As Richard says, 'But it is certainly a warm comfort.'

This chapter is where you'll start building Bayesian models. Bayesian models learn from prior information and this is super helpful in the small world. If this assumptions are close to reality, then they are also great in the large world.

2.1.1 Garden of forking paths

This is the humble beginnings of Bayesian inference: counting and comparing possibilities. Richard compares this to Jorge Luis Borges' short story "The Garden of Forking Paths." In short, life is full of paths and exploring all of them will help make good inference. As we learn. We prune. This inference might not give a correct answer in the large world but it can guarantee the best possible answer in small world, given the information fed to it.

2.1.1.1 Counting possibilities

Here we'll take a play on the marble scenario but use fish in a bucket. Let's say you have a stream with either Bull Trout (*salvelinus confluentus*) or Rainbow Trout (Oncorhynchus Mykiss) and let's say there are 4 samples from that stream. There are five different possibilities in these samples: some with all and some mixed,

```
## # A tibble: 4 x 5
##      p1    p2    p3    p4    p5
##   <dbl> <int> <int> <int> <dbl>
## 1     0     1     1     1     1
## 2     0     0     1     1     1
## 3     0     0     0     1     1
## 4     0     0     0     0     1
```

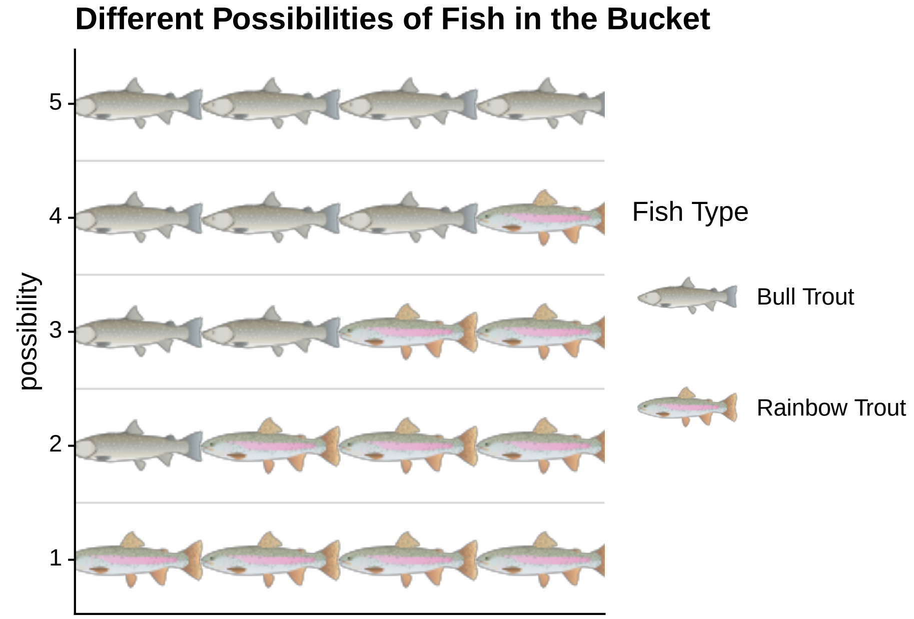


Figure 2.1: Possibilities in the bucket.

We want to figure out what conjecture is most plausible, given the evidence about the samples. Let's say that in the first 3 samples (with replacement) you get: Bull Trout, Rainbow Trout, Bull Trout (BT, RB, BT). Then what is the most plausible configuration in the bucket? Let's consider one example; `p2` (possibility #2) from Figure 2.1 (BT, RB, RB, RB). This is where the forking paths comes in handy because we can count the ways that this possibility can happen (Figure 2.2), i.e. possibility meaning there is 3 RB and 1 BT in the bucket.

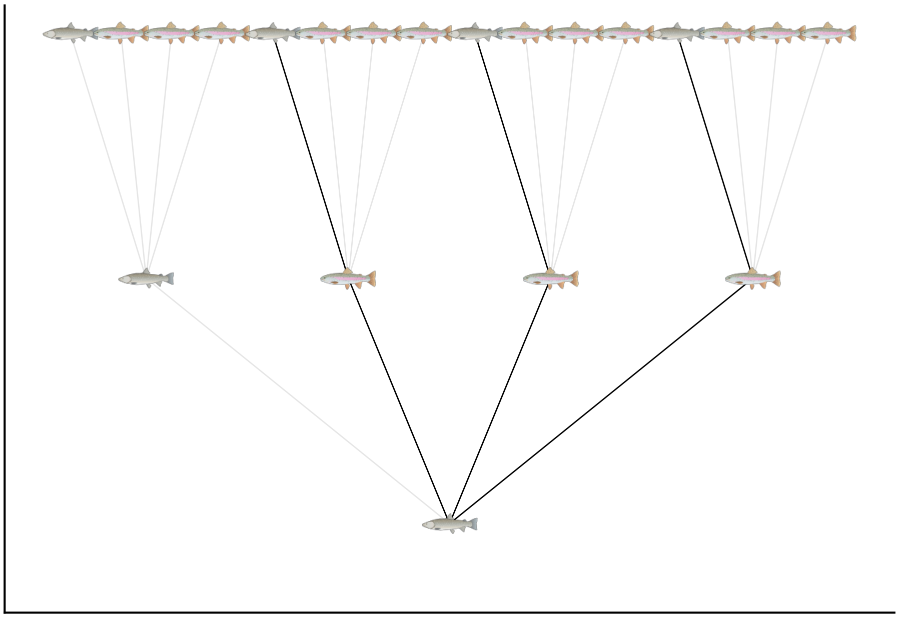


Figure 2.2: Paths taken in the bucket of forking fish.

In Figure 2.2 above, we can see that there is 1 way of getting a bull trout in the first draw, 3 ways in the second and 1 again in the third. We can now do this for all the possibilities. Figure 2.2 shows that there are 3 *paths* or *ways* for the bag to contain (BT, RB, RB, RB). Now this is great but we need to test all of the ways right? For example, let's say that we wanted to see if the bucket had all RB!? Right, that doesn't make sense. We pulled BT twice so this definitely isn't possible. Same goes for all BT. Since these paths don't exist based on our reality, we know that the way to produce them is 0. Now that we know the ways to produce (BT, RB, RB, RB) and the 0's, let's look at the others in Table 2.1 below.

We can see based on our sample that (RB, BT, BT, BT) has 9 ways and (RB, RB, BT, BT) has 8 ways. These 5 different conjectures give us different *paths* through the garden of forking data and tell us how we could produce (BT, RB, BT). As you can see in Table 2.1, we just multiply the intermediary steps instead of counting but you'll see later that this is really handy when the conjectures grow.

Table 2.1: Conjectures and Ways to Produce Fish in a Bucket

p_1	p_2	p_3	p_4	draw 1: Bull Trout	draw 2: Rainbow Trout	draw 3: Bull Trout	ways
BT	BT	BT	BT	4	0	4	
RB	BT	BT	BT	3	1	3	
RB	RB	BT	BT	2	2	2	
RB	RB	RB	BT	1	3	1	
RB	RB	RB	RB	0	4	0	

Table 2.2: Updating priors with new data.

p_1	p_2	p_3	p_4	Ways to Produce: Bull Trout	Prior counts	New count
BT	BT	BT	BT	4	0	0
RB	BT	BT	BT	3	9	27
RB	RB	BT	BT	2	8	16
RB	RB	RB	BT	1	3	3
RB	RB	RB	RB	0	0	0

2.1.1.2 Combining other information

What if we had information about the relative plausibility of each conjecture? Many ways we might have this but the important point is that it can help us to update these conjectures. In our case there's an easy solution: just multiply the counts.

Let's consider our initial data and call them *priors*. Now let's say we draw another fish out of the bucket and it's a Bull Trout (BT). We could just start all over again and count the paths, etc or we could just multiply by the new observation and ways to produce it given the conjecture. It turns out these methods are mathematically identical (independent assumption). So to do this, we'll just take the *prior* counts based on (BT, RB, BT) <=> (0, 3, 8, 9, 0) and multiply by the new ways (0, 1, 2, 3, 4) (Table 2.2).

This is great because as new data arrives we can just update!

2.1.1.3 From counts to probability

This

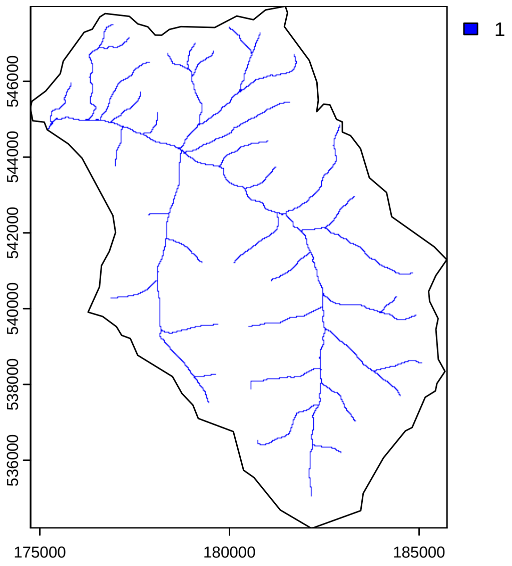
2.1.2 Building a Model

The globe problem always reminded me of a similar hydrology problem with predicting headwater streams. We can look at the streams across the landscape similar to how we look at a tossing globe. First we need to sample a *theoretical* stream network across a landscape and just like the globe if we get an observation on a stream or not we should be able to generate a posterior distribution. So let's bring in some data. We'll look at the NHDPLus High Resolution raster in Northwest Montana and sample from this raster to get an idea for a proportion of land to water.

```
library(terra)

nhd <- rast('Z:/GIT/stat_rethinking/stat_rethinking_2023/nhdStrOrdRast.tif')
sutton <- vect('Z:/GIT/stat_rethinking/stat_rethinking_2023/sutton.shp') %>% project(crs(nhd))
nhd_sutton <- crop(nhd, sutton, mask = T)

plot(nhd_sutton, col = 'blue')
plot(sutton, add = TRUE, fill = NA)
```



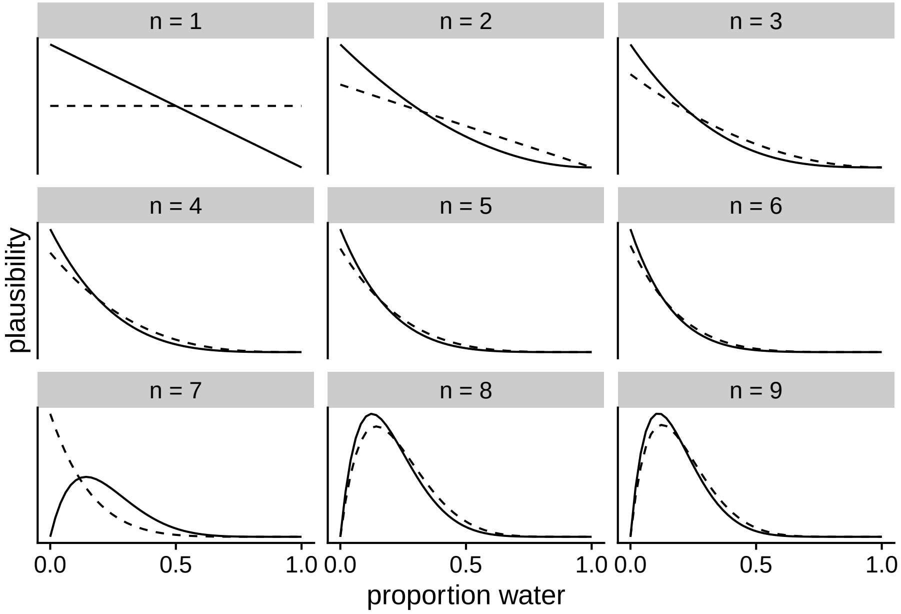
```
set.seed(1234)

sample9 <- terra::spatSample(nhd_sutton, size = 9, method = "random", replace = TRUE)

## just adding a S and L
sample9$observations <- ifelse(is.na(sample9$nhdStrOrdRast), 'Land', 'Stream')
```

In this example we'll randomly sample this stream raster to get a sample of S (stream) or L (land). So from the vector `sample9` we get Land, Land, Land, Land, Land, Land, Stream, Land, Land.

As you can see it's pretty rare to get a stream with all that land!



So we can see from the graph above that we are taking a 50% probability of getting 'S' or 'L' (even though we know that's not true…) and then updating based on our sample of Land and Stream. So every time `Land` is observed, the peak of the plausibility curve moves to the left in the graph. Now that we've seen the model we can just use the `dbinom()` to find the probability.

```r
dbinom(1, size = 9, prob = .5)
```

```
## [1] 0.01757813
```

This is the relative number of ways to get 1 stream, holding p at 0.5 and $N = W + L$ at 9. We can now summarize our model.

$$W \sim \text{Binomial}(N, p)$$

And the unobserved parameter p gets:

$$p \sim \text{Uniform}(0, 1)$$

This is not the best prior since we know this area has way more land to stream.

2.1.3 Making the model go

A Bayesian model can update all of the prior distributions to their purely logical consequences: the Posteriro Distribution. This distribution contains the relative plausibility of different parameter values, conditional on the data and model, e.g. probability of the parameters, conditional on the data $Pr(p|W, S)$. This means the probability of each possible value of p, conditional on the specific S and L that we observed.

Bayes' theorem:

$$Pr(S, L, p) = Pr(S, L|p)Pr(p)$$

Now just reverse the p on the rhs.

$$Pr(S, L, p) = Pr(p|S, L)Pr(S, L)$$

Not that these are equal to the same thing we can just equal them to each other.

$$Pr(p|S, L) = \frac{Pr(S, L|p)Pr(p)}{Pr(S, L)}$$

$$\text{Posterior} = \frac{\text{Probability of the data} \times \text{Prior}}{\text{Average probability of the data}}$$

Chapter 3

Sampling the Imaginary

We describe our methods in this chapter.

Math can be added in body using usual syntax like this

3.1 math example

p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt(\frac{p(1-p)}{n}) \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$

You can also use math in footnotes like this[1].

We will approximate standard error to 0.027[2]

[1] where we mention $p = \frac{a}{b}$

[2] p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt(\frac{p(1-p)}{n}) \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$