

# Basic Models

Rachael Caelie (Rocky) Aikens

6/10/2020

## A Simple Model with Diagnosis by Severity

Suppose we want to do a study of a disease. Perhaps our target population is “all people,” or perhaps it is some subset of people (e.g. only Men, only Asians, etc.). Each person,  $i$ , has an underlying (unobserved) disease state  $Z_i$  and a disease severity  $S_i$ . A simple model is:

$$\begin{aligned} Z_i &\sim_{iid} \text{Bernoulli}(0.15) \\ S_i | Z_i = 1 &\sim_{iid} \text{Uniform}(0, 1) \end{aligned}$$

Where  $S_i = 0$  whenever  $Z_i = 0$ .

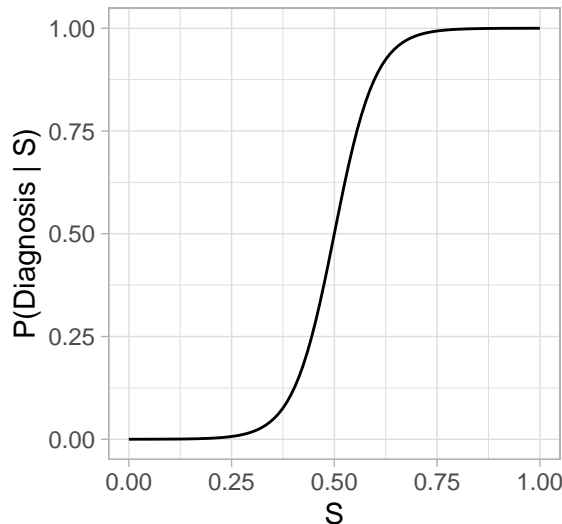
Suppose additionally that a person’s likelihood of diagnosis is a function of the severity of their illness, namely

$$D_i \sim_{iid} \text{Bernoulli}(\phi(S_i))$$

$\phi(S_i)$  could be many possible functions, here I use a sigmoid function:

$$\phi(S_i) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 S_i))} - c$$

Here,  $c = \frac{1}{1 + \exp(-\beta_0)}$  is a corrective constant to ensure that  $\phi(0) = 0$ . Let  $\beta_1 = 20$  and  $\beta_0 = -10$ . This gives the following probability of diagnosis as a function of severity.



A few example rows sampled from the above distribution are printed below:

##	disease	severity	p_diagnose	diagnosed
## 1	0	0.0000000	0.0000000	0
## 2	0	0.0000000	0.0000000	0
## 3	0	0.0000000	0.0000000	0
## 4	0	0.0000000	0.0000000	0
## 5	0	0.0000000	0.0000000	0
## 6	1	0.8705389	0.9993503	1

## A Naive Study

Let's suppose that a naive researcher would like to study this disease. In particular, they want to ask

1. What is the prevalence of the disease?
2. What is the distribution of severities of this disease?

This researcher collects  $5 \times 10^4$  records from the EHR,  $\{S_i, D_i\}_{i=1}^n$ . We'll assume for the moment that the individuals represented in the EHR are representative of the target population (this is almost certainly false - it's inevitable that there are types of people that systematically tend not to appear in the EHR at all). We will suppose that not all information about all of these individuals is recorded (i.e. the underlying disease states for all individuals are not known).

Because they observe the diagnoses but not the underlying disease states, the naive researcher considers only the diagnosed people to have the disease.

### Prevalence

Out of  $5 \times 10^4$  people, 3781 are diagnosed, so they estimate the prevalence to be 0.07562. Because they collected so many samples, they are quite sure of this number. The table below shows the confidence intervals that this researcher might calculate based on their approach.

method	x	n	mean	lower	upper
asymptotic	3781	50000	0.076	0.073	0.078

### Severity Distrubution

Now they want to understand the severity of the disease. Let's suppose for the moment that we have a straightforward measure of disease severity between 0 and 1. The researcher calls any severity less than 1/3 "low" severity, 1/3 - 2/3 "medium" severity and 2/3-1 "high" severity. From these counts, they conclude that this disease more frequently occurs at high severities, and that mild forms of the disease are rare.

severity_grp	count	prevalence
low	14	0.00
medium	1256	0.33
high	2511	0.66

Finally, they want an estimate of median severity and quantiles. They find that the median severity is quite high, based on the severity quantiles in the diagnosed people, below

	x
0%	0.0851674
25%	0.6204654
50%	0.7535650
75%	0.8746784
100%	0.9998469

The researcher concludes that this is a disease that appears in about 7.56% of the population and that when it appears it tends to be severe.

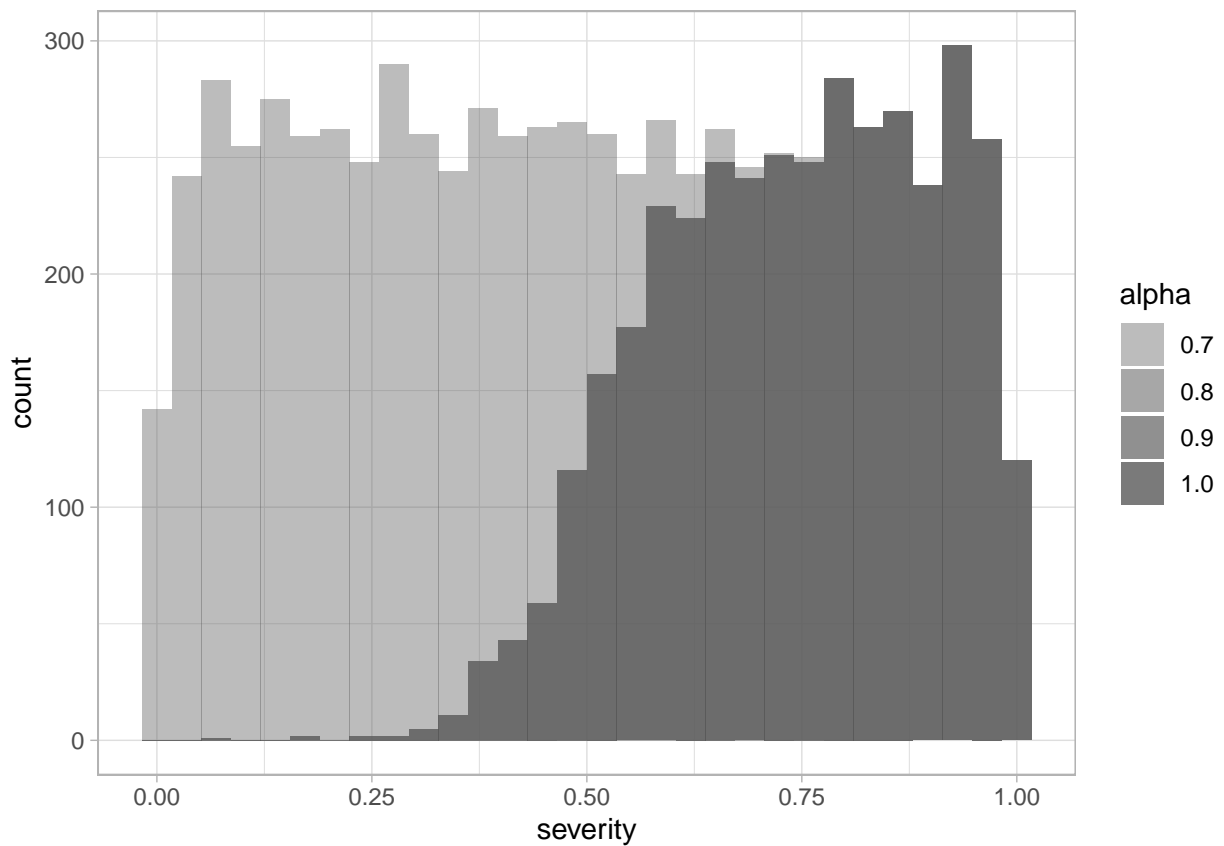
## What's wrong with this approach

The issue here is that the researcher acts as though everyone who has the disease is diagnosed. In their calculations of prevalence they are estimating  $P(D_i = 1)$ , but they imagine they are estimating  $P(Z_i = 1)$ . In their calculations of severity they imagine they are understanding  $P(S_i)$ , when in fact they are estimating  $P(S_i|D_i = 1)$ .

Under this model, we have a disease prevalence of 15% and a diagnosis rate of about 50%. Already we can see how misdiagnosis can distort our perception of the disease - If we consider only the people being diagnosed for the disease, the disease prevalence appears to be about half what it actually is.

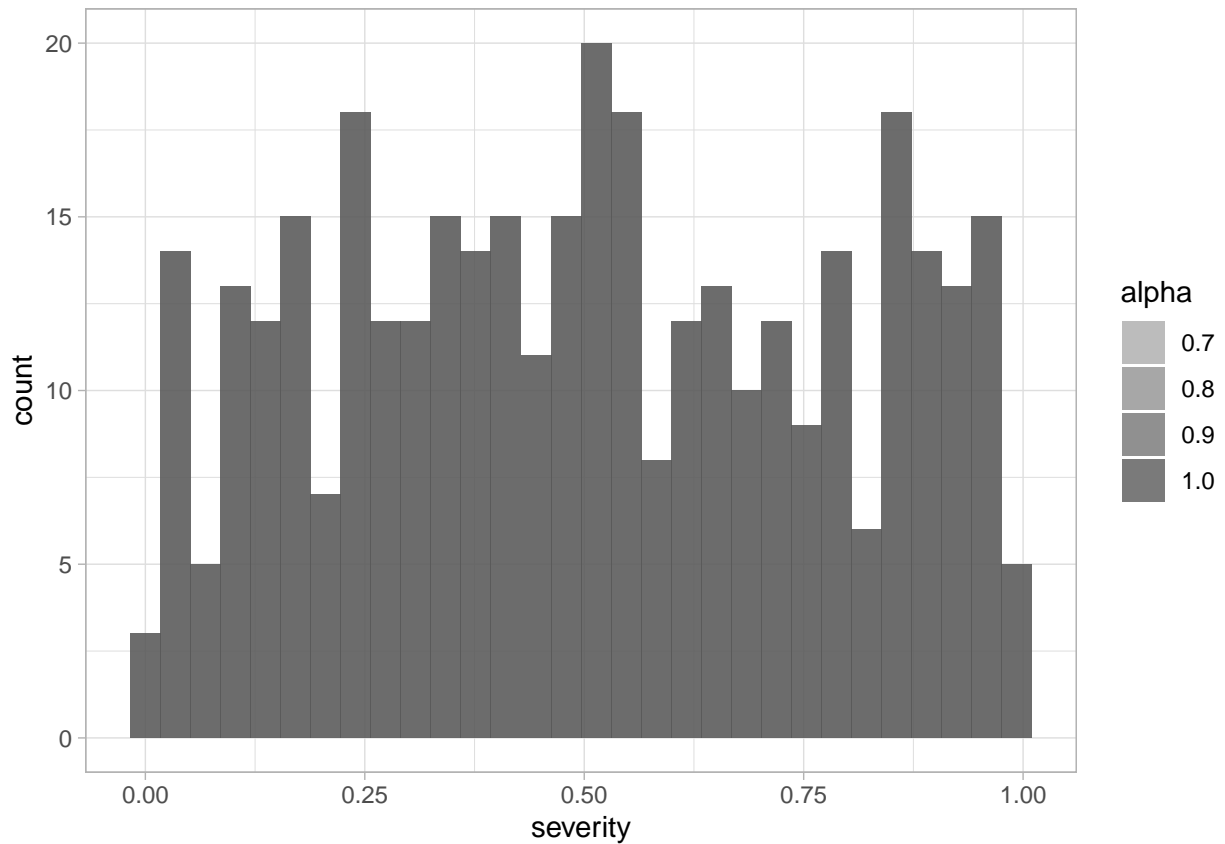
disease	Total	Diagnosed
0	42429	0
1	7571	3781

This diagnosis process distorts the distribution of severities we observe.



## What Happens When we Sample

Suppose up until now that every diagnosis has been made with a diagnostic test, which we'll assume for now has perfect sensitivity and specificity. Now suppose that for 5% of the population we decide to test them anyway, regardless of the severity of their symptoms, we retrieve the uniform distribution in the randomly tested sample.



From this random sample we can obtain a new estimate of disease prevalence (with wider confidence intervals now because of the smaller sample)

method	x	n	mean	lower	upper
asymptotic	368	2495	0.147	0.134	0.161

	Actual	Estimate	Lower CI	Upper CI
Total Cases	7571	7375	6679	8070
Missed Cases	3790	3594	2898	4289

## Some Maths

Here's some maths that might be useful...?  $i$  subscripts are dropped for simplicity

$$P(D = 1|S) = \frac{P(S|D = 1)P(D)}{P(S)}$$

It seems like all the quantities on the right hand side are knowable quantities, and it would be nice to understand the left hand side so we could understand how this distorts our measurement process. Moreover

$$P(S) = P(S|Z = 1)P(Z = 1) + P(S|Z = 0)P(Z = 0)$$

I'm less clear on how this would be useful to us, but the left hand side could be a knowable quantity and the right hand side is all things it would be nice to know. If we can reasonably assume that  $P(S|Z = 0) = 0$  (i.e. the symptoms/severity in question cannot be explained by any other disease), then the problem simplifies substantially. Then we could estimate the number of symptomatic cases by identifying all people with  $S \neq 0$ , and the problem reduces to identifying the cases where  $Z = 1$  but  $S = 0$ .

Also

$$P(Z) = \frac{P(DZ)}{P(D|Z)} = \frac{P(Z|D)P(D)}{P(D|Z)}$$

Which seems useful? Because we might be able to get a handle on  $P(Z|D)$  from our understanding of whatever diagnostic criteria we use, and  $P(D)$  is estimable from the data.

## Things that could be changed about this set-up

Even at this simple stage there are tweaks that could be made. Here are some:

1. We could vary the distribution of severities  $S_i|Z_i = 1$ . For example these could be normal (most people medium ill), right tailed (most people mildly ill), or left-tailed (most people very ill). There is no reason the severity must be between zero and 1.
2. We could vary the probability of diagnosis based on severity,  $\phi(S_i)$
3. Rather than considering severity, we could consider some number of discrete symptoms,  $W_{i1}, \dots, W_{ik}$ , and have  $\phi(W_{i1}, \dots, W_{ik})$ .