Why Outlier Analysis?

The Height Example.

Picture this: late one afternoon, a distant friend texts you out of the blue, trying to convince you that they are a totally ordinary height. You haven’t seen them in years and can’t remember their height, so you’re interested in what your friend has to say. They tell you that they grew up in an average neighborhood and their parents who were of average height. They never had an especially nutritious diet, but rarely got sick, either. In other words, an ostensibly typical childhood.

*Okay, sure — so how tall are you?* is the natural question on your lips. The answer? *7 feet.* Of course, you laugh, certain that they’re lying to you.

But why are you so sure?

For one thing, you probably don’t know anyone who is 7 feet tall. You might have a few friends who are around 6 foot 3, but most of the people you’ve met are somewhere between 5 foot 4 and 5 foot 11, with a few people around 5 feet. In other words, if you were to plot how many people you know at each height, the picture would look something like this:

*[insert approx. normal distribution of heights]*

Even though you don’t know it, that picture (called a histogram), is floating around somewhere in your subconscious. When your friend tells you they’re 7 feet tall, alarm bells start ringing — *I’ve never met anyone who is 7 feet tall, so there is a very low chance that my friend is telling the truth….*

Your brain had put that numerical answer **into context** to decide how to interpret it, ultimately making you suspicious. But if you didn’t know anything about how tall most people are, you would have no idea what to make of your friend’s answer.

Gerrymandering

Gerrymandering is drawing political boundary lines with an ulterior motive. While it comes in many flavors, we’ll focus on partisan gerrymandering, where one political party controls the process and tries to exaggerate its own political dominance. The party in power does this in two ways: packing, where they seek to stuff their opponents into districts with very high percentages, and cracking, where they disperse their opponents into several districts in numbers too small to predominate. Both of these techniques generate wasted votes by the opponents and thus reduce the opponents’ share of seats relative to their share of the votes cast. [cite Moon]

Here’s a quick example:

*[5x5 grid example, split 40/60, showing how you could get 100% seats for the majority depending on how you draw the lines]*

One of the only constraints redistricters have to follow when drawing boundary lines is to keep roughly the same number of people in each district. Above, each districting plan has 5 voters per district. However, in one plan 100% of the representatives are Party A, while in another plan only 60% of the representatives are. It all depends on how the lines are drawn.

[explain more about seats]

Why should you care about gerrymandering? In a representative democracy, your vote is your voice — but if your vote is separated from other like-minded citizens or packed together with tons of like-minded citizens, your voice will either be drowned out or redundant.

*[example of packing / cracking]*

Partisan Gerrymandering Metrics [EG example — but I could do MM or other].

Many political scientists, mathematicians, and lawyers have put forth formulas that claim to measure how much a state has been gerrymandered. These formulas, or **metrics**, take in the districting plan of a certain state and output a number. Take one famous partisan metric, for example: the Efficiency Gap.

Clarify that you could use any metric/measurement, or seats histogram…

The Efficiency Gap actually takes two inputs: the districting plan for a certain state, and voting data from a specific election. Since gerrymanderers seek to maximize the number of wasted voters for their opponent, EG tries to measure the difference between the wasted votes for each party.

*[insert efficiency gap equation]*

What actually counts as a “wasted vote?” The way the EG defines it, in any given district, all the votes for the losing party count as wasted, since the outcome would have been the same had each of those voters decided not to go to the ballot box that day. Moreover, every vote more than 50% for the party that did win *also* counts as wasted, since that party would have won anyways if those voters had stayed home.

EG can be positive, negative, or zero. If it’s zero, that means the number of wasted votes is the same for each party — ostensibly, no party has been trying to pack or crack the other. The more negative or positive it is, the more evidence there is that one party is trying to waste the votes of its opponents.

Try it out yourself! Color in the districts of the state below — once you’re finished, click the “Calculate Efficiency Gap” button and see what you find.

*[Districtr embedding, along with a “Calculate EG” button that prints the EG for the drawn districting plan]*

How do you feel about this answer: EG = xx?

You probably are not as incredulous as when your friend told you they were 7 feet tall. In contrast to your natural knowledge of the distribution of people’s heights, you have no idea which EG’s are common and which are rare. In other words, you have no sense of where EG = xx lies on the histogram.

To make this histogram, you’ll need a lot of examples of districting plans (just like you needed to know many people’s heights in order to understand how rare 7 feet is). Rather than drawing thousands of plans by hand, you can use a computer algorithm called **MCMC** to sample the plans and build a histogram. Try it out below!

*[insert simplified dashboard like I’ve drawn out before, set on the same state as above, with number of steps, one or two constraints, and a ‘run’ button. After ‘run’ is clicked, animations will play showing the progression of plans on the state map as well as the EG histogram being built, with EG = xx shown for context as a vertical red line. For increased comprehension, the first few steps should be kinda slow, then speed up so that the whole ~10,000 step thang runs kinda quick?]*

Now you can see how your districting plan stacks up compared to other districting plans. There are **way more** than 10,000 possible plans, but this test gives you a pretty quick and easy way to judge your EG **in context** of other ones — exactly the same way you judged that your friend was lying to you by comparing their stated height to other heights you knew.

This type of thinking is called **outlier analysis**. The EG of a districting plan on its own cannot tell you if gerrymandering is going on, but if you look at it compared to all other possible values in that state and see that it is an outlier, you have strong reason to be suspicious.

How much of an outlier is too much? This is a tough question because there is no one simple answer. If your friend had told you they were 6 foot 6 inches, would you have believed them? What about 6 foot 5? 6 foot 2? Outlier analysis doesn’t tell you which plan you *should* pick. In fact, it doesn’t always tell you which ones you shouldn’t pick. Take the example below — if you are trying to consider whether a plan with EG = yy is gerrymandered and it falls at the red line in the histogram, you may not be able to conclusively say it’s an outlier.

*[histogram with something that isn’t at the median but isn’t definitively an outlier]*

However, outlier analysis can sometimes tell you when something is clearly an outlier, as in the case of these graphs:

*[real histograms of egregious gerrymandering as shown by EG]*

The above plots are EG histograms with the EGs of X, Y, Z state’s real districting plans as comparison. Outlier analysis shows that these are super unlikely to occur just by chance, making you suspicious — just like a friend who claims to be 7 feet tall.

How do we know that we sampled from the right distribution? Why is a random sample the right thing to compare it to?

Further Questions

Now that you have a sense of an EG histogram, can you use it to provide context for the Efficiency Gap for any state, for any election? Unfortunately, no. Specific characteristics of the political situation in each state end up affecting EG in different ways, so you’ll have to run MCMC sampling on each state to see what that state’s histogram looks like, given certain election data. Think of the heights example — below are examples of heights histograms for Sweden and India — see the difference?

*[insert a histogram with two sets of data, colored different ways to show that the median for Indian heights is way lower than the median for Swedish heights]*

If your friend is from India and they tell you they are 7 feet tall, you’d be more suspicious of them than if they were from Sweden, since many more Swedes are taller! EGs work in the same way. [expand on this probably]

[more questions?]