Grounded Stats and Better Plots

And How to Drill Down to a Take-Away Message

Bliss Cohen

March 10, 2021

Background

Per your favorable comments about Cassie Kozyrkov, I have read some of her articles and watched a few videos. I am now a huge fan! I think I fully converted after watching her video using puppies to explain p-values *(Why are p-values like needles? It's dangerous to share them!)*. Anyway, these articles have added more context to what you have been telling me about statistics:

- Don't waste your time on statistics
- Data Science's Most Misunderstood Hero
- Statistics for people in a hurry
- Never start with a hypothesis
- Statistical inference in one sentence
- A trick question for data science buffs
- Why are p-values like needles? It's dangerous to share them!
- What is correlation? Not causation.
- What is the difference between analytics and statistics?

Now to test my understanding. I believe this overly simplistic situation would involve "pure" statistics:

A business is considering a product improvement launch to enhance user's comfort. It is necessary to carefully design a study with the right power to determine the appropriate action:

- Don't launch (default action)
- Launch (alternative action)

A study would then test these hypotheses:

- The product improvement is not more comfortable (null hypothesis)
- The product improvement is more comfortable (alternative hypothesis)

In the real world, however, I am not in the position of designing studies nor am I making decisions. So far, I have been using statistics as an exploratory tool to "validate" something I feel is worth highlighting. In other words, the statistics give me the confidence to call something out. Some of Cassie's comments really resonated with me.

> "For example, statisticians might forget the equations for a t-test's p-value because they get it by hitting run on a software package, but they never forget how and when to use one, as well as the correct philosophical interpretation of the results. Analysts, on the other hand, aren't looking to interpret. They're after a view into the shape of a gory, huge, multidimensional dataset. By

> knowing the way the equation for the p-value slices their dataset, they can form a reverse view of what the patterns in original dataset must have been to produce the number they saw. Without an appreciation of the math, you don't get that view. Unlike a statistician, though, they don't care if the t-test is right for the data. They care that the t-test gives them a useful view of what's going on in the current dataset. The distinction is subtle, but it's important." *Data Science's Most Misunderstood Hero*

> "When you're doing data work just for yourself, think about why you're doing it and what will change depending on what you see. If you have a clear answer, you'll also see the decisions. If you're truly not making decisions, then the data science subfield you want is analytics, not statistics. Analytics may use hypothesis testing equations, but does not actually test hypotheses. (Just like using a scalpel doesn't mean you're doing surgery.)" *A trick question for data science buffs*

> "If you're interested in analytics (and not statistics), p-values can be a useful way to summarize your data and iterate on your search. Please don't interpret them as a statistician would. They don't mean anything except there's a pattern in these data. Statisticians and analysts may come to blows if they don't realize that analytics is about what's in the data (only!) while statistics is about what's beyond the data." *Why are p-values like needles? It's dangerous to share them!*

Plots (and tables) highlight key points. But at this juncture, if anything was a little muddy in my brain it now gets even muddier. I vacillate between providing too much information or not enough. No doubt part of the problem is that I haven't clearly stated the plot's purpose, even to myself.

I want to use statistics appropriately for my role. I also want to make better plots and learn when a table trumps a plot. After making a point, I would like to follow-up with some kind of "recommendation".

As an aside, eventually I want to develop a logo for myself. But before I can develop a logo, I need to deeply understand and articulate what I offer. I feel like nuggets to this answer lie somewhere in Cassie's articles.

Objective

Discuss the following:

- How to use statistics appropriately
- How to improve plots
- How to frame "recommendations"

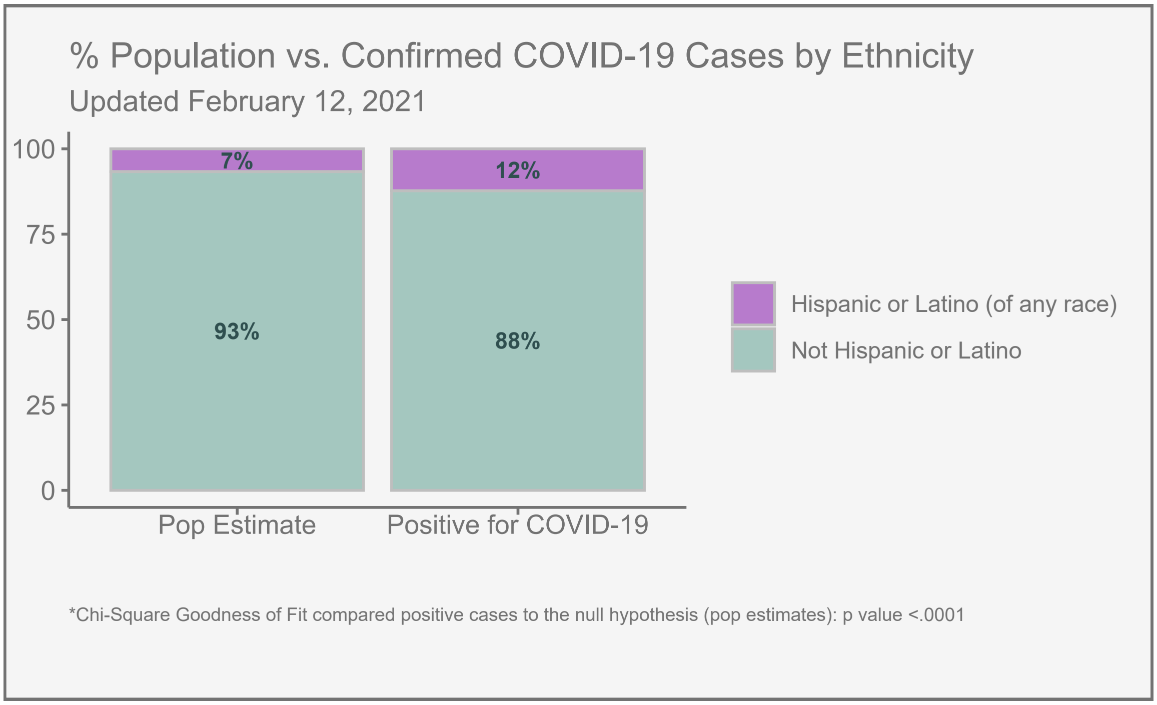
The 3 sections below set the stage for the Questions at the end.

1. Get Real on Statistics

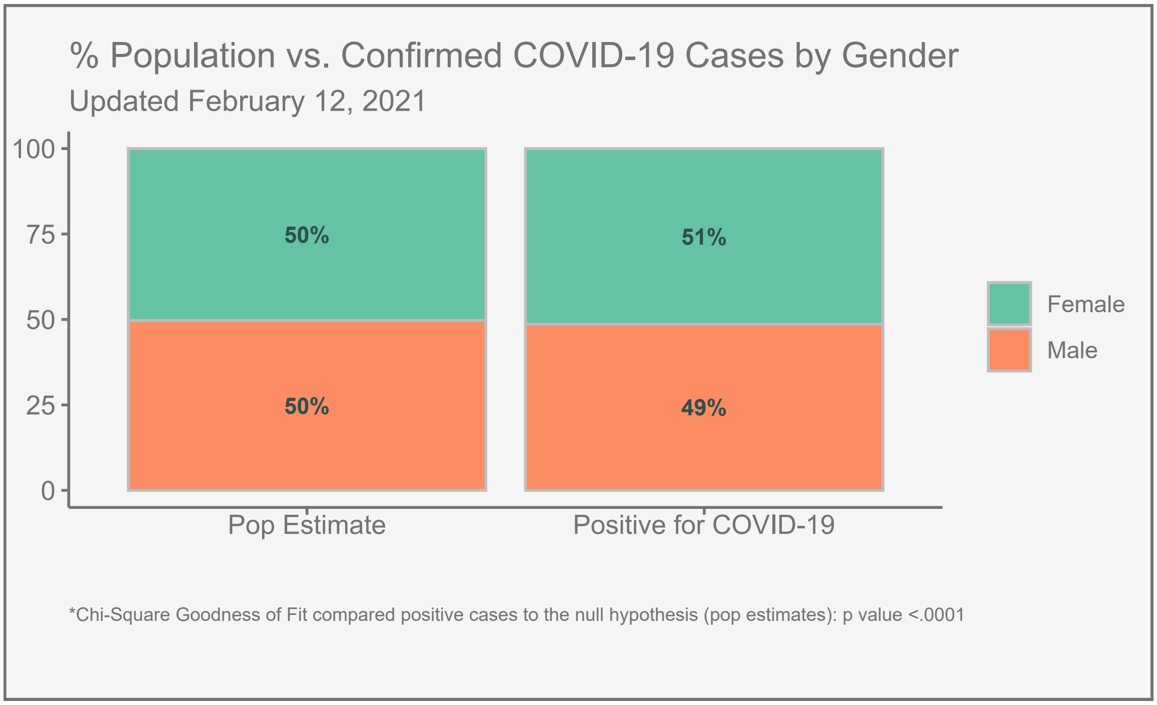
I have been pulling raw data from the Wisconsin's health department to compare COVID-19 infections among different demographics. I conducted Chi-Square Goodness of Fit to compare infections against population estimates to assess potential skews.

Demographic plot purpose: Highlight where demographic disparities exist

Our eyes and the statistics tell us that Hispanics or Latinos are more likely to test positive for COVID-19 than Non Hispanics or Latinos.



On the other hand, even though Gender differences were significantly different - women were more likely to test positive than men - the difference doesn't seem meaningful. Is 51% infected women really different than 50% women in the population?

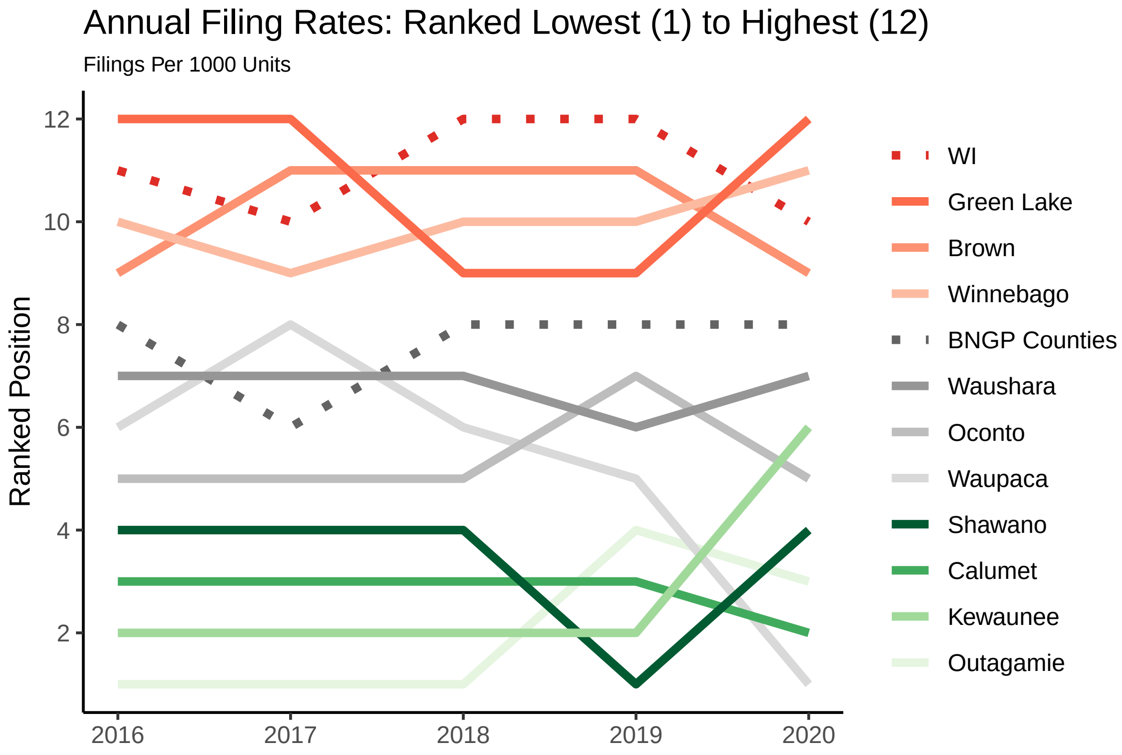


2. Did I Move the Needle on this Plot?

You've seen the first two plots below in Evictions_Part4.Rmd.

An annual eviction filing rate was calculated for various regions by dividing the number of filings by the number of occupied rental units. The rate was ultimately expressed as the number of filings per 1000 rental units (filings/unit*1000).

Filing rates were then ranked from 1 to 12 with 1 being the lowest and 12 being the highest. The low, medium, and high groups are visualized below across the years.

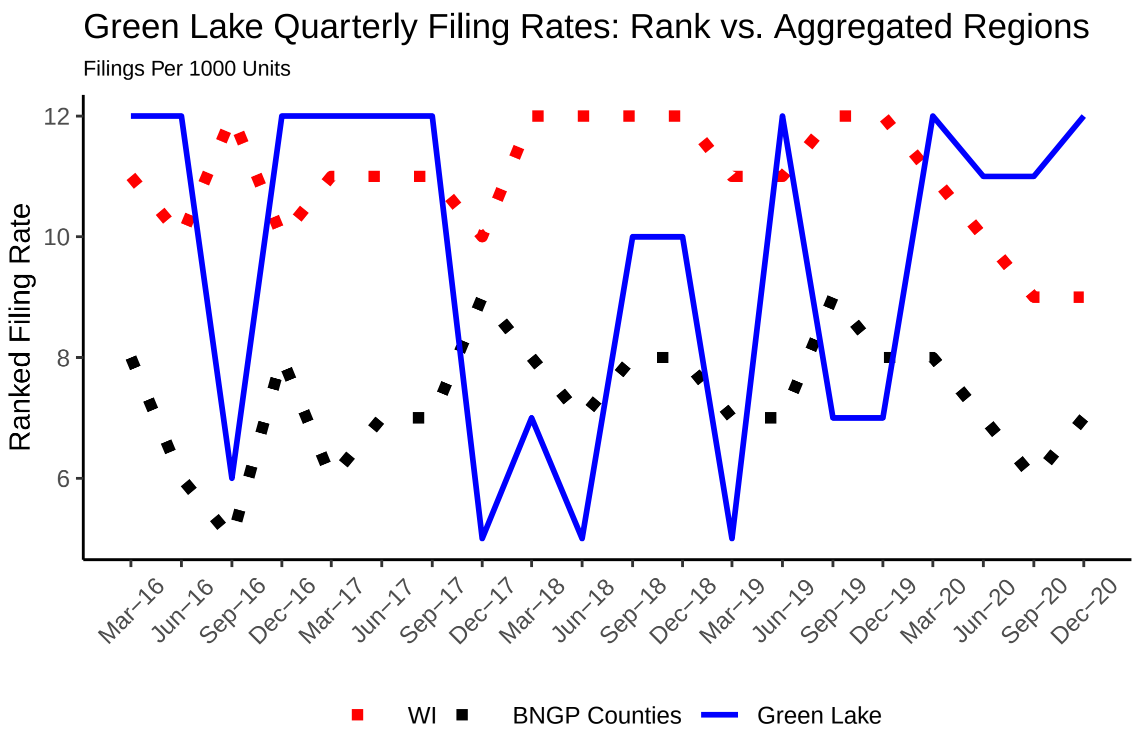


Year aggregations provide a 30,000 foot view. Quarterly aggregations allow for a more ‘real time’ snapshot.

Quarterly plot purpose: Offer an exploratory tool to investigate how a region’s ranking might be changing during 2021

The original plot below was an attempt to compare a single county with aggregated regions per quarter. I wanted to make it clear that a non-aggregated region (county) was being compared to aggregated regions, so I didn’t include other counties. Like the yearly ranked plot above, I used dashed lines to identify aggregated regions (logic: add parts and pieces to make a whole).

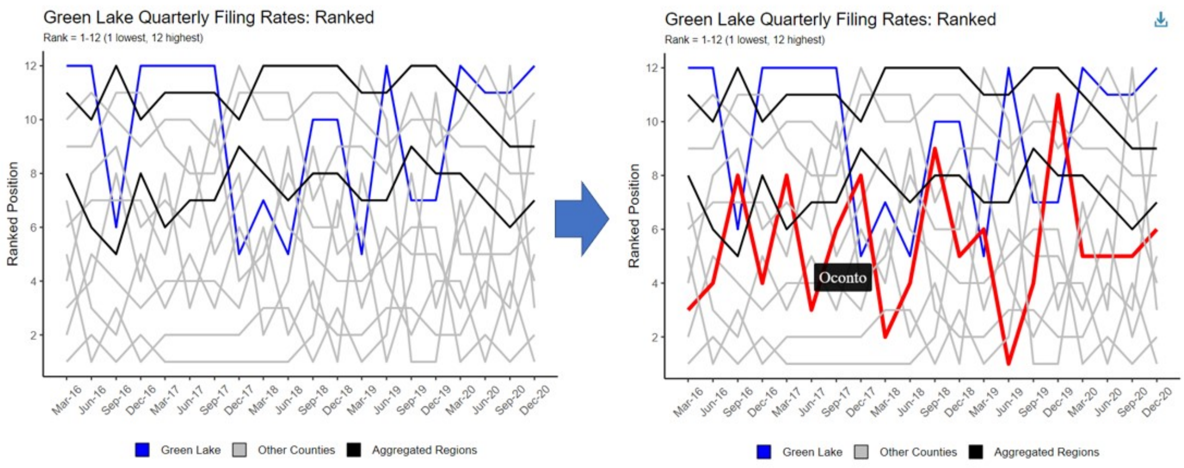
OK, I realize the graph shows quarters between 2016-2020. Obviously, we haven’t finished 1Q2021. The idea would be to maybe include the second half of 2020 plus 2021 quarters. I’m showing the historical record because that’s what I have.



I now see that the above plot is confusing. I revised my approach to include all regions in one plot. The following plots compare the same county of interest - Green Lake - to **ALL** other regions.

Interactive graphs can't be rendered in pdf, so I am resorting to low-resolution screen shots to give a feel for how they might look in a Shiny app, especially when the user hovers over a line.

Again, the idea would be to start with the second half of 2020 and add 2021 as the quarters accrued.

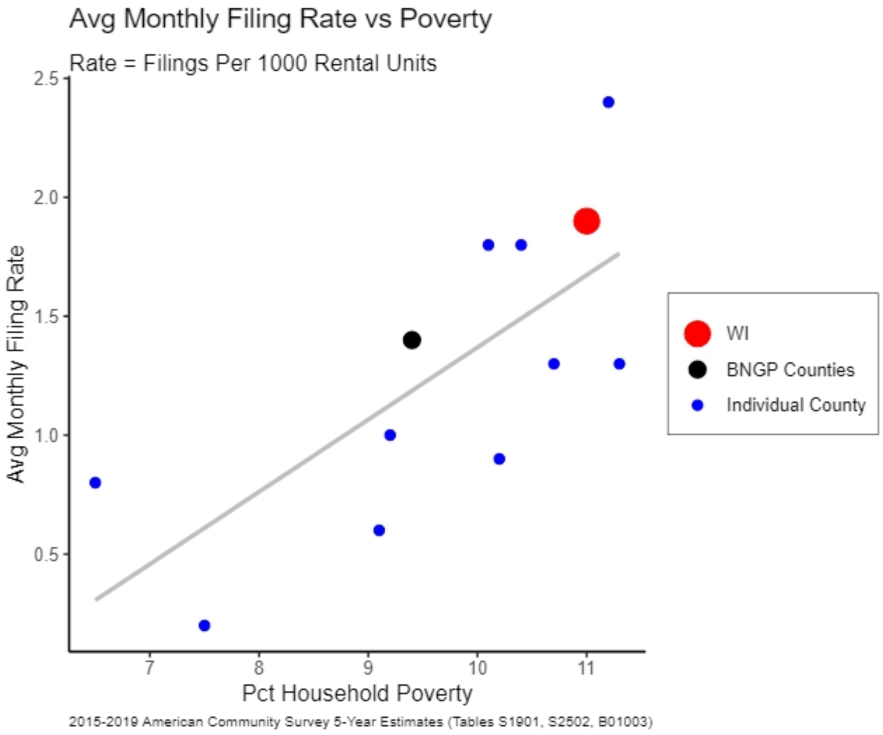


3. And What About Recommendations?

In Evictions_Part4.Rmd, I used multiple linear regression to examine the average yearly filing rate against the following features:

- `Avg HH Income`
  - Average Household Income
- `Pct Poverty HH`
  - Percent households below Official Poverty Level
- `Pct ALICE HH`
  - Percent households above the Official Poverty Level but unable to meet basic needs
- `Units Per 1000 People`
  - Number of occupied rental units per 1000 people

Except for 2020, the `Avg Yearly Filing Rate` correlated with `Pct Poverty HH`; regions with higher % poverty were more likely to have higher eviction filing rates. Poverty is an extreme situation, and eviction is an extreme action, so this association is not terribly surprising.



2015-2019 American Community Survey 5-Year Estimates (Tables S1901, S2502, B01003)

I understand that you can't say "reduce poverty and then you will reduce eviction filings". But is there anything you can say within the context of a recommendation? Would it be fair to say something like this:

"Any initiative designed to reduce eviction filings may want to consider programs that reduce poverty"

Questions

1. Am I thinking about the role of statistics correctly?

2. Should I have bothered to include the statistical comment in the caption on the demographic graphs (type of test and p-value)? My thought is “no”. I see a pattern, I believe the pattern - including the p-value makes it feel like I’m hiding behind it.

3. Was the revised quarterly ranked plot (shown via interactive screen shots) any better than the original? In other words, is it better to compare a region to all other regions, making it clear in the legend which ones are aggregated?

4. Within Cassie’s framework, I definitely fall into the Analytics world rather than Machine Learning or Statistics. Machine Learning, and especially Statistics, feel closer to making decisions than Analytics. But is there a place for an analyst’s recommendations? Or should recommendations be framed in terms of a question?

5. How do you reconcile the word “hypothesis” in these different contexts? On the one hand, Cassie says Don’t use the word hypothesis when you’re doing analytics, or you’ll sound like an idiot. On the other hand she says Analytics helps you form hypotheses. It improves the quality of your questions

6. This article, Three Common Hypothesis Tests All Data Scientists Should Know, preaches in a very typical way. What would Cassie say about this article?

7. I found Cassie’s references to “Frequentist” and “Bayesian” intriguing. I am familiar with the “Frequentist” language, but not “Bayesian”. Are you aware of any dataset that would be a good introduction to Bayesian, just so I can start learning about it? Is the Smartphone and Motion project we started way back when a “Bayesian” problem?

Btw. . . I enjoyed reading your Data Science Method articles. The visualization section in “Exploratory data analysis” touched on some of the stuff we’ve been talking about. The Microsoft Bing story was really interesting. Although the phrase “personal bugbear” really threw me - that’s a new one!☻

Anyway, reviewing your articles made me think about my struggles with visualization. Kieran Healy’s book, Data Visualization: A Practical Introduction, is a good reference. Chapter 1 dissects various types of visualization “badness”. The rest of the book shows how to use ggplot to make a huge range of plots. Although the book is helpful, I fantasize about some kind of “Visualization of the Week” with a discussion of what is good and what isn’t and how it can be improved. Just one visual snippet at a time without a whole lot of other stuff. Are you aware of any exercises like this?