Spatial Aggregations

CSC544

Consider the lowly histogram

Exploring Histograms, an essay by Aran Lunzer and Amelia McNamara

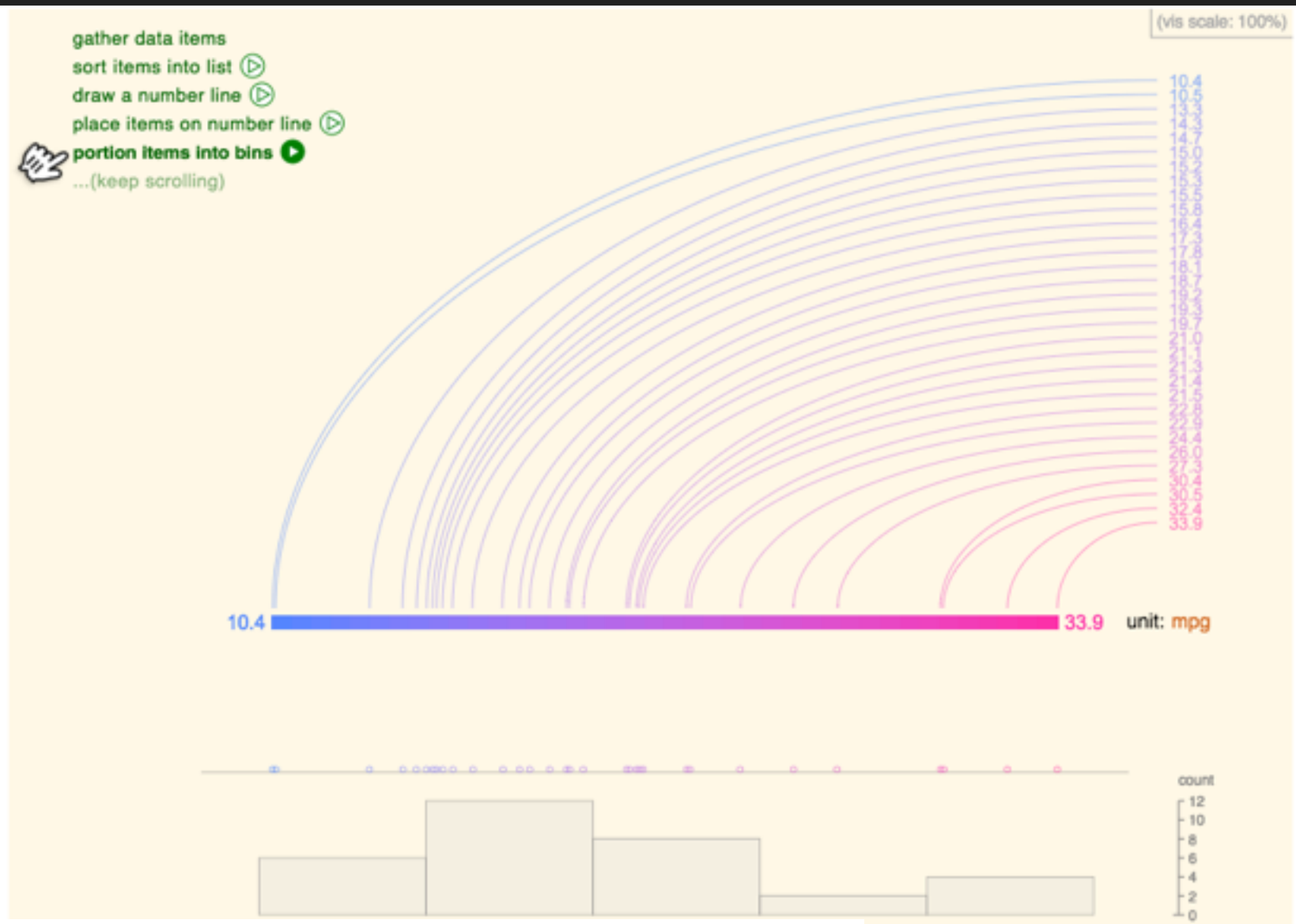
Portioning items into bins—the essence of a histogram

Once items are placed along a number line, drawing a histogram involves sectioning the number line into **bins** and **counting** the items that fall into each bin. Notice how the distribution shown in the histogram echoes the distribution from the dot plot.

Gathering the items into bins helps us to answer the question "what is the distribution of this data like?" Imagine trying to describe some dataset over the phone: rather than mechanically reading out the entire list of values, it would be more useful to provide a summary, such as by saying whether the variable's distribution is symmetric, where it is centered, and whether it has extreme values. A histogram is another kind of summary, in which you communicate the overall properties in terms of portions (i.e., bins) of the data.

For example, the "Geyser" data can be described as being bimodal (because its histogram has two 'peaks'), while "NBA" is more unimodal, and perhaps right-skewed (because the bin heights decrease towards the right).

Maybe because histograms are visually similar to bar charts, it's easy to think that they are also similarly objective. But, unlike bar charts, histograms are governed by many parameters. Before describing a dataset to someone based on what you see in its histogram, you need to know whether different parameter values might have led you to different descriptions.



Bin-breaks: Why these bins?

For a start, you probably noticed that the histograms shown for our sample datasets have different numbers of bins. This is because we used Sturges' formula, a common method for

http://tinlizzie.org/histograms/

models

Consider the lowly histogram

- But how do we choose those free parameters (bin offset, bin width) ?

- Bin width: “mostly solved”

  - Many similar results: some constant times “variability” divided by the cube root of the number of points

  - Freedman-Diaconis: $$\text{Bin size} = 2\,\frac{\text{IQR}(x)}{n^{1/3}}$$

- Freedman-Diaconis: $$\text{Bin size} = 2\,\frac{\text{IQR}(x)}{n^{1/3}}$$

Consider the lowly histogram

- What about offset? AFAIK, nothing known

- BUT
  - what happens if we take the average of all possible choices?

Kernel Density Estimation

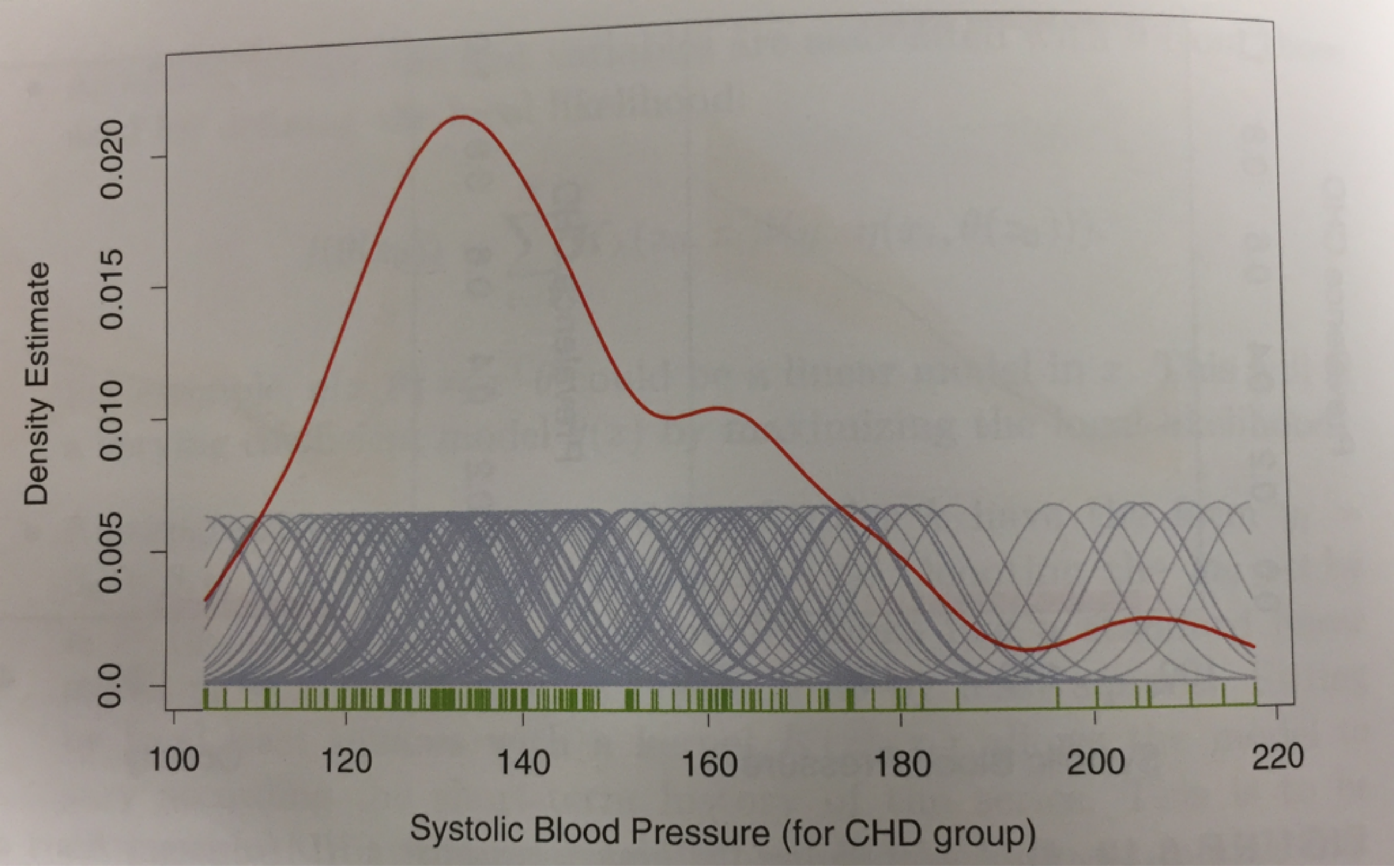
$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)$$

- (What?)

Kernel Density Estimation

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)$$

- (What?)



Hastie et al.,
Elements of Statistical Learning

Kernel Density Estimation

- But, how do we choose that free parameter?
- Many similar results: some constant times “variability” divided by the **fifth** root of the number of points
- Silverman’s rule of thumb:

$$h = \left(\frac{4\hat{\sigma}^5}{3n} \right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},$$

Histograms vs KDEs

- How do we compute each?

- Why is this still not completely a solved issue?
  - Can you notice a problem with the bin width rules?

Choropleths

Choropleth

https://www.nytimes.com/interactive/2014/11/04/upshot/senate-maps.html

https://codepen.io/sassquad/pen/qZRaOd

Choropleth

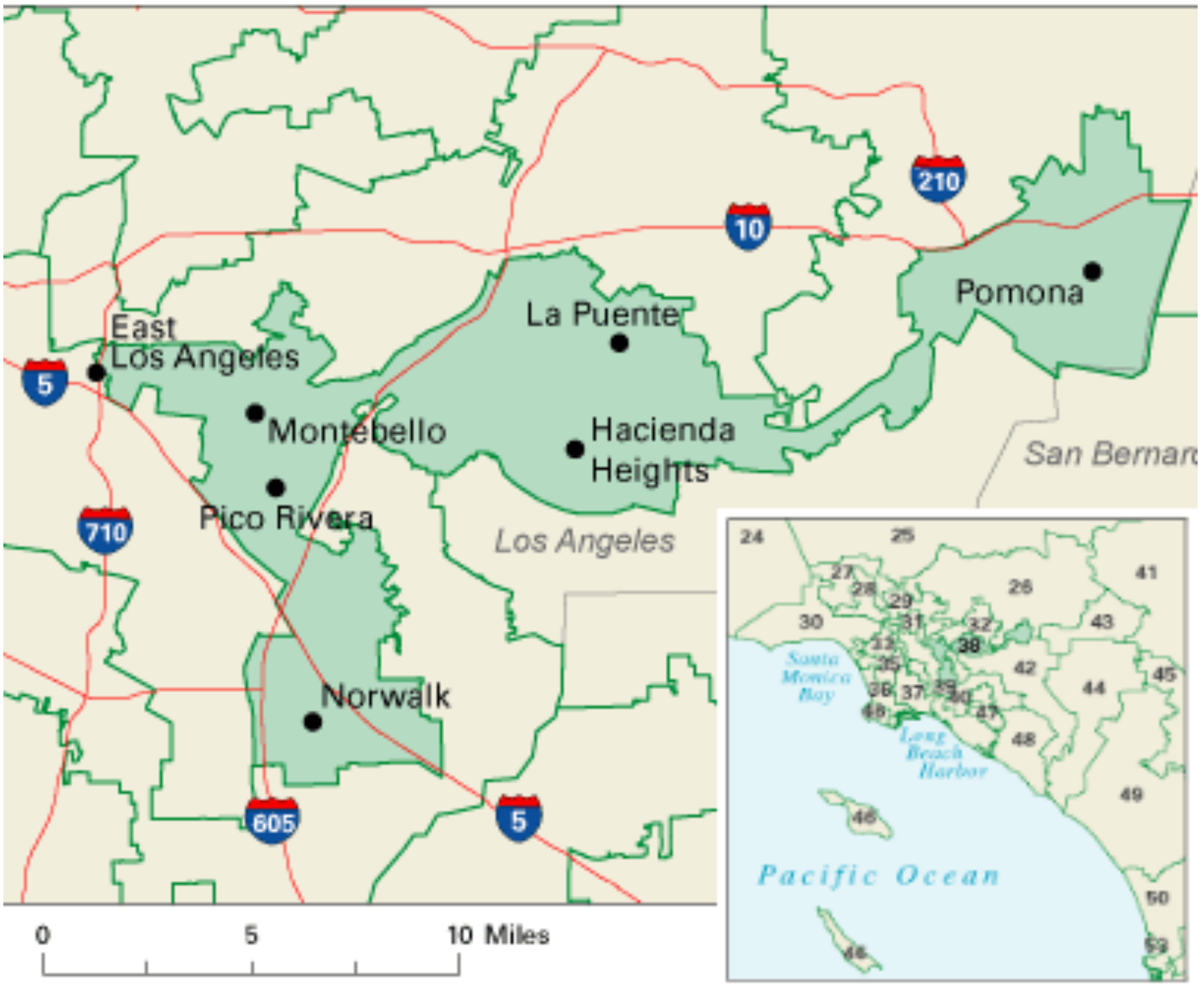
- What are the perceptual risks?

- What are the statistical risks?

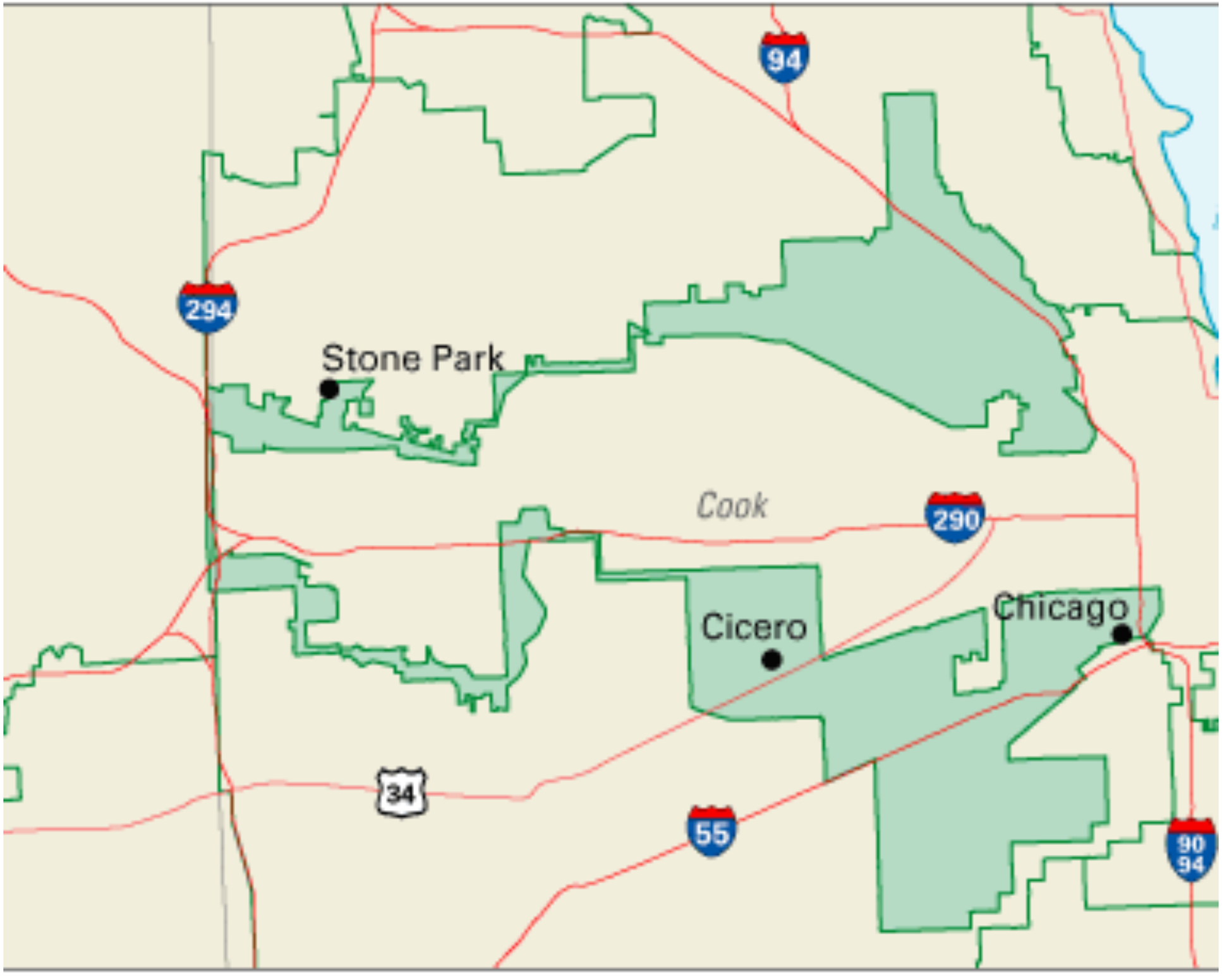
Choropleth

- MAUP: the “modifiable areal unit problem”
  - How did you split that region?

Congressional District 38



Congressional District 4



Heatmap

https://demographics.virginia.edu/DotMap/

Histogram : KDE ::

Choropleth : Heatmap