Jenks Natural Breaks

Instructor: Wei Ding

What is "Natural Breaks"?

- "Natural breaks" finds the "best" way to split up the ranges.
- Suppose we had 30 counties, 15 counties with 0-1 values, 10 counties with 16-18 values, and 5 counties with 24-29 values. Obviously, the "best" ranges are 0-1, 16-18, 24-29. Counties with similar values should have the same color. "Natural breaks" is the only method that finds the "best" ranges.

What do we mean by "best ranges"? -

- It means the ranges where like areas are grouped together.
- Obviously, we do not give a low rate area the same color as a high rate area. Natural breaks minimizes the variation within each color, so the areas within each color are as close as possible in value to each other.

How does natural breaks work?

- Natural breaks is complicated because many steps are involved, but does not involve any higher math.
- Step #1 is simple - Calculate the "sum of squared deviations for array mean" (SDAM).
- Assume four counties, with values: 4, 5, 9, 10.
- Mean = 7.
- SDAM = $(4-7)^2 + (5-7)^2 + (9-7)^2 + (10-7)^2 = 9 + 4 + 4 + 9 = 26$.

How does natural breaks work?

- Step #2 is complex - For each range combination, calculate "sum of squared deviations for class means" (SDCM_ALL), and find the smallest one.
- SDCM_ALL is similar to SDAM, but uses class means and deviations. Suppose we have four counties and two ranges.
- For [4][5,9,10], SDCM_ALL = $(4-4)^2 + (5-8)^2 + (9-8)^2 + (10-8)^2 = 0 + 9 + 1 + 4 = 14$.
- For [4,5][9,10], SDCM_ALL = $(4-4.5)^2 + (5-4.5)^2 + (9-9.5)^2 + (10-9.5)^2 = 0.25 + 0.25 + 0.25 + 0.25 = 1$.
- For [4,5,9][10], SDCM_ALL = $(4-6)^2 + (5-6)^2 + (9-6)^2 + (10-10)^2 = 4 + 1 + 9 + 0 = 14$.
- [4,5][9,10] has the smallest SDCM_ALL, so is "best ranges", minimizes variation within classes. Intuitively, it makes sense to use [4,5][9,10], and the natural breaks algorithm automatically figures this out.

How does natural breaks work?

- Step #3 is simple - As a final summary measure, calculate a "goodness of variance fit" (GVF), defined as (SDAM - SCDM) / SDAM.
- GVF ranges from 1 (perfect fit) to 0 (awful fit). Higher SDCM_ALL (more variation within classes) results in lower GVF.
- In the examples in step #2, GVF is (26 - 1) / 26 = 25 / 26 = 0.96 for the best combination, and (26 - 14) / 26 = 12 / 26 = 0.46 for the two rejected combinations, a huge difference.