# **Data Description**

The U.S. state of California has been experiencing issues with water availability. One culprit related to this shortage is the fact that the majority of the water for the state comes from the northern third of the state while the majority of the demand is in the southern two-thirds of the state.

Seemingly, the shortage of water supply in California would also be exacerbated by more water usage. Dr. Heyman obtained a dataset that summarizes the water usage across all California counties in the years 1995, 2005, 2010, and 2015. Specifically, the total per capita water usage of the public water supply (in gallons/person/day) is the type of usage that you’ll examine.

(Original Data Source: <https://waterdata.usgs.gov/ca/nwis> )

(If you are unfamiliar with the county structure within U.S. states, the Wikipedia link below shows California counties. <https://en.wikipedia.org/wiki/List_of_counties_in_California> )

# **Questions**

We are interested in whether there is evidence that the average total per capita water usage of the water supply (across all counties in California) differs for 1995 to 2015. Although the data are adjusted for differing populations across the counties already, the location of the county within the state may be related to the public water supply available for usage. But, we’re not specifically interested in whether the average usage differs for at least one county, since there are several intuitive reasons why county per capita usage would differ on average.

1. Explain why we are using a statistical hypothesis test for population means instead of just directly comparing sample means.

Suppose that your professor suggested using a paired test for two population means to compare the average total per capital water usage of California public water supply usage in 1995 and 2015.

1. What variable would your professor be suggesting to use to pair observations/cases across?

Unfortunately, it would be quite unreasonable to assume that errors are independent to use a paired test for two population means.

1. Specific to *this* scenario/dataset, what a main source of dependence that is remaining? Why?

# **Solutions**

1. Notice that the provided data are just public water supply usage. There are other water supplies available that were used during these years as well. Thus, we have information regarding part of the population. (It is questionable whether the public water supply usage is representative of all water supply usage.)
2. Observations/cases would be paired across county.
3. A main source of dependence that is not accounted for would be dependence among counties. Counties that are proximate may have similar water supply usage. The description even states that the location of the county may be related to the amount of water available for use, giving the answer here.