In Lecture #3 from this week, we provided an overview of non-probability sampling and some of the limitations of this type of approach for generating data and then making inferences about larger populations based on the data. While non-probability sampling methods can generate a lot of data very quickly and at low cost, analysts of the data must be very careful when making population inferences based on the data. There are many pitfalls to making these larger conclusions based on data generated using this technique, and here we consider a rather spectacular failure of this approach for trying to make statements about larger populations.

In 2008, researchers studied the potential of analyzing Google searches to try and understand the spatial and temporal distribution of a flu epidemic. The researchers wanted to see if the analysis of Google searches could effectively replicate the conclusions that CDC researchers were finding based on analyses of data from formal probability samples of the U.S. population. At first, the findings seemed to be almost in exact alignment with the CDC data, and this was viewed as a tremendous success of "big data". However, these findings were based on a limited window of time. As people continued to study the Google data relative to CDC data, substantial differences emerged, and the Google data was largely viewed as providing a misleading picture of the spread of the flu virus during that time period. This study received a great deal of coverage in the popular press: for more details, take a look at this article: https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/.

The point of this example is not to say that all non-probability samples will lead to erroneous conclusions. When carefully and properly applying one of the two inferential approaches introduced in Lecture #3, one can make sound conclusions about the features of a larger population. In another very popular example, Wang and colleagues analyzed hundreds of thousands of survey responses from Xbox users in 2012, and used a type of calibration weighting approach to make conclusions about voting intentions of the population. Their estimates of preference for presidential candidates in the 2012 election were almost spot-on with forecasts based on aggregation of polling data. For more on this study, please see this link: https://www.sciencedirect.com/science/article/pii/S0169207014000879?via%3Dihub.

In short, it is always important to first ask what type of sampling mechanism was used to generate the data set that is presently under consideration. Then, if non-probability sampling methods were used, careful analyses of how representative that sample is with respect to the target population of

interest are necessary before proceeding. We will consider analytic techniques for these types of data in more detail in future lectures.

Additional Deep-Dive Readings on Non-Probability Sampling

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J., and Tourangeau, R. (2013). Report of the AAPOR Task Force on Non-Probability Sampling. American Association for Public Opinion Research, May 2013. Available from www.aapor.org.

Elliott, M.R. and Valliant, R. (2017). Inference for Non-Probability Samples. Statistical Science, 32(2), 249-264.

Pasek, J. (2015). When Will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence. International Journal of Public Opinion Research. doi:10.1093/ijpor/edv016.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2014). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.