One post that has attracted my attention recently is a post from The Upshot with title “How One 19-Year-Old Illinois Man Is Distorting National Polling Averages”.

Data journalists from New York Times looked on how polls reported by different agencies may differ. There will always be differences in polls but what seemed to be odd about polls from U.S.C./LAT was that it consequently reported 5 points below, compared to others. The poll seemed to be biased somehow.

They were able to acquire the data and reengineer the polls. It turned out the cause of the difference in polls was pinpointed to one 19-year-old black man in Illinois. He was sure to vote for Donald Trump.

The polls used weights to compensate for categories where data is particularly hard to gather. For example if the data gathered in group young man between 18-29 contains 100 observations and 3 of them vote for Donald Trump then one way to weight the broad category would be to assume that 3% of votes in that group would go to Donald Trump. But if there are only 10 observations and 1 of them vote for Donald Trump is it then 10% in that category as a whole? Well, it’s hard to say and some assumptions are more risky than others.

It seems that U.S.C./LAT made a couple of unusual decisions that produced some odd results. The only time the poll was close to other polls, with more typical weights, was when the 19-year did not participate in the poll at all.

I think that the explanation is plausible. It points to the fact that gathering data from some groups has become difficult, particularly for young people. One part of the explanation could be that traditionally data was gathered by using land line phones. That is no longer an option for young people today. And pretty hard for most of the population since everybody mostly uses cell phones.

It is also clear that polls have harder times to come up with reliable results. Good examples are polls for Brexit and Colombian Referendum that both failed to predict the outcome.