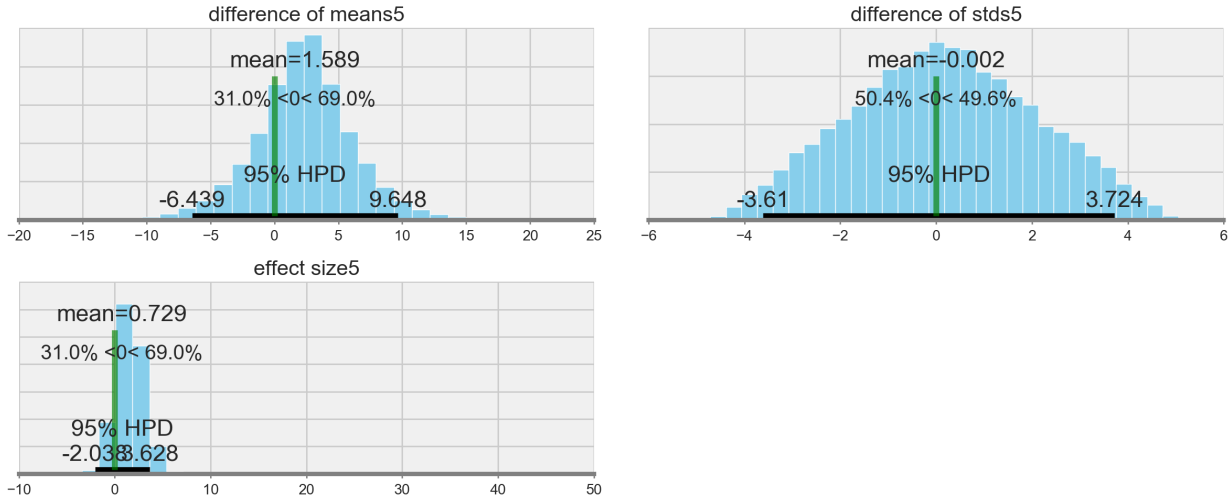
J Beightol
1/19/2016

In this analysis I'll be discussing the results from my exploration of the START global terrorism dataset. After exploring the data I decided upon which populations I would focus on for my Bayesian test.

Bombings were the most prolific type of terror attack, and bombings they were the most deadly type of attack as measured by the amount of deaths per attempted attack. While looking at stats of the most terror prone areas in the world, I noticed that India and Pakistan both experienced many attacks, but India had significantly more bombings compared to its neighbor to the north. Given the parallels between the two countries, I thought it would be interesting to see if terror attacks in India were deadlier than attacks in Pakistan. One other consideration was the time period to study, since START collection methodology has varied from 1970 to the present day, I decided to focus on a specific time period. I chose the 1990s as those were a volatile period in Indian-Pakistani relations.

Once I made my choice I had to choose a Bayesian prior for both the populations. I ended up opting for a conservative prior: I would set the prior to be deaths per attack for all South Asian countries in the 1990s. I could have selected specific priors for each country, such as the rate of deaths per attack for the 5 years preceding, and the 5 years following the decade for each country. However by selecting this conservative prior I ensured that any difference I found would be immune to criticism. I could be sure of my findings.



The average amount of deaths from terror attacks across South Asia was 3.4, so I set my range of standard deviations to be 0 to 6. After running 10,000 simulations with this data, I found that the mean attack in India killed 3.429 individuals and the mean attack in Pakistan took the lives of 1.84 individuals. Ultimately, the differences between these means were not enough to declare

them to be different populations that had a meaningful difference between their average death toll from terrorist attacks.

Next I moved towards imputing the values for worldwide bombings in 1993. I debated a few different methods, and ultimately decided to use the 1993 data located in the appendix of the code book including the # of incidents, the deaths, and the injuries for 1993. I had to decide which years I wanted to consider to train my model. Ultimately I opted for using the 3 years before, and the 3 years after 1993. While more data could have been helpful, I worried that as we got further away from 1993, the additional data might have a negative effect on my model.

First I built a decision tree regressor, which enabled me to see how the decision trees were using the X values to predict the amount of bombings. Unsurprisingly, my decision tree put a lot of weight into the number of terrorist events in order to determine the number of bombings.

My main concern with the output data from my decision tree is that my model could have been too overfit to the limited years surrounding 1993. In order to address this concern, I built a random forest regressor which considered the same inputs as my random forest, but gave me a slightly lower results.

The random forest regressor is an amalgamation of many different decision trees, which helps protect the model from over fitting.

Ultimately I averaged the results of my two models together, and got a final value 1685 bombings in 1993.