Investigating the Quality of FAOSTAT, OIE, Census, and National Data

Hypothesis

1. If you take all the growth rates for a population of animals in a country every year, and you plot every yearly growth rate, it should produce a normal curve. From there we can spot outliers by defining a cut off of 3 standard deviations. Past this point we can label those points as outliers and conclude that there is a high chance that they are inaccuracies.
   1. Great idea!
2. If there is more than one source of data for a population of animals in a given country for a specified year or time range, we could find the average population between the sources. Depending on how far off from that average a source is, we can conclude there is an increased chance that the data is wrong
   1. Not sure about taking the average – I like the idea of observing the difference between sources but I think that there are issues and differences with collection strategies that would mean that for some countries and some species over a particular period of time, their numbers would be closer to the “actual” number. Instead, it might be interesting to look at the difference between the sources, breaking it up by species and maybe looking at specific periods of time (5-10 year intervals). Like you did with hypothesis 1, I think that we might measure the +- differences (maybe as a percentage) of difference sources against FAOSTAT (or the national data) to see if there are consistent patterns, e.g. are WOAH counts always smaller or do national figures track nicely with FAOSTAT for cattle but not for pigs, etc.
3. We could take the average rate of change for the increase in a species population for each year for a given country for each data source. We could compare the rate of change between each data source and if one is substantially bigger than the other two it could be concluded that there is a higher chance that it is wrong.
   1. Love this idea!
4. Using the interquartile range: We would find the average ROC for the growth in animal populations. Then find the interquartile range (<https://www.scribbr.com/statistics/interquartile-range/>) for the data. Then add error bars around the data to visualize where approximate outliers are.
   1. I think that this would be very interesting to see – it would make for a dashboard that would help members of GBADs evaluate the confidence that they have in the data that they are using. I will be very interested to see this visualization.

