On Rankings Econ 20150, Principles of Statistics Kevin R Foster, CCNY Spring 2012

We often see statistics reported that rank a number of different units based on a number of different measures of outcomes. For instance, these could be the US News ranking of colleges, or magazine rankings of city livability, or sports rankings of college teams, or any of a multitude of different things. We would hope that statistics could provide some guidance; we would hope in vain.

In the simplest case, if there is just a single measured variable, we can rank units based on this single measure, however even in this case there is rarely a clear way of specifying which rankings are based on differences that are large and which are small. (The statistical theory is based on "order statistics.") If the outcome measure has, for example, a normal distribution, then there will be large number of units with outcomes right around the middle, so even small measurement errors can make a big difference to ranking.

In the more complicated (and more common) case, we have a variety of measures of outcomes and want to rank units based on some amalgamation of these outcomes. A case where a large number of inputs generates a single unit output looks like a utility function from micro theory: I face a choice of hundreds (or thousands) of different goods, which I put into a single ranking: I say that the utility of some bundle of goods is higher than the utility of some other bundle and so would rank it higher (even if both were affordable).

However there is no way to generate a composite utility function that completely and successfully takes account of the information of individual choices! (This result is due to CCNY alumnus and Nobel Laureate Ken Arrow.)

Many rankings might take an equal weighting of each item, but there is absolutely no good reason to do this generally: why would we believe that each measure is equally valid? Some rankings might arbitrarily choose weights, or take a separate survey to find weights (equally problematic!). You could average what fraction of measures achieve some hurdle.

One possible way around this problem is to just ask for people's rankings (let them figure out what weights to use in their own utility functions) and report some aggregation. However here again there is no single method that is guaranteed to give correct aggregations. Some surveys ask people to rank units from 1-20, then add the rankings and the unit with the lowest number wins. But what if some people rank number 1 as far ahead of all of their competitors, while others see the top 3 as tight together? This distance information is omitted from the rankings. Some surveys might, instead, give 10 points for a #1 ranking, 8 points for #2, and so on – but again this presupposes some distance between the ranks.

This is not to say that ranking is hopeless or never informative, just that there is no single path that will inerrantly give the correct result. Working through the rankings, an analyst might determine that a broad swathe of weights upon the various measures would all give similar rankings to certain outliers. It would be useful to know that a particular unit is almost always ranked near the top while some other one is nearly always at the bottom.

Examples:

Sports might seem to have it relatively easy since there is a single ranking given by prearranged rules, but still fans can argue: a team has a good offense because they scored a lot (even though some other team won more games); some players are better on defense but worse on offense. Sports Illustrated tried to rank the 100 all-time best sports stars, somehow comparing baseball player Babe Ruth with the race horse Secretariat! Most magazines know that rankings drive sales and give buzz.

College rankings try to combine student/faculty ratios, measures of selectivity, SAT scores, GPA; some add in numbers of bars near campus or the prestige of journals in which faculty publish. What is best?

Food nutrition trades off calories, fat content, fiber, vitamin and mineral content; who is to say whether kale or blueberries are healthier? Someone trying to lose weight would make a very different ranking than someone training for a marathon.

Sustainability or "green" rankings are difficult: there are so many trade-offs! If we care about global warming then we look at CO₂ emissions, but what about other pollutants? Is nuclear power better than natural gas? Ethical consumption might also consider the material conditions of workers (fair-trade coffee or no-sweatshop clothing) or other considerations. Politics: which political party is better for the economy? Could measure stock returns or unemployment rate or GDP growth or hundreds of others. Average wage or median earnings (household or individual)? Each set of measures could give different results.

Other Ignorant Beliefs

While I'm working to extirpate popular heresies, let me address another one, which is particularly common when the Olympics roll around: the extraordinary belief that outliers can give useful information about the average value. We hear these judgments all of the time: some country wins an unusual number of Olympic medals, thus the entire population of the country must be unusually skilled at this task. Or some gender/race/ethnicity is overrepresented in a certain profession thus that gender/race/ethnicity is more skilled on average. Or a school has a large number of winners of national competitions, thus the average is higher.

Statistically speaking, the extreme values of a distribution depend on many parameters such as the higher moments. If I have two distributions with the exact same mean, standard deviation, and skewness, but different values of kurtosis, then one distribution will systematically have higher extremes (by definition of kurtosis).