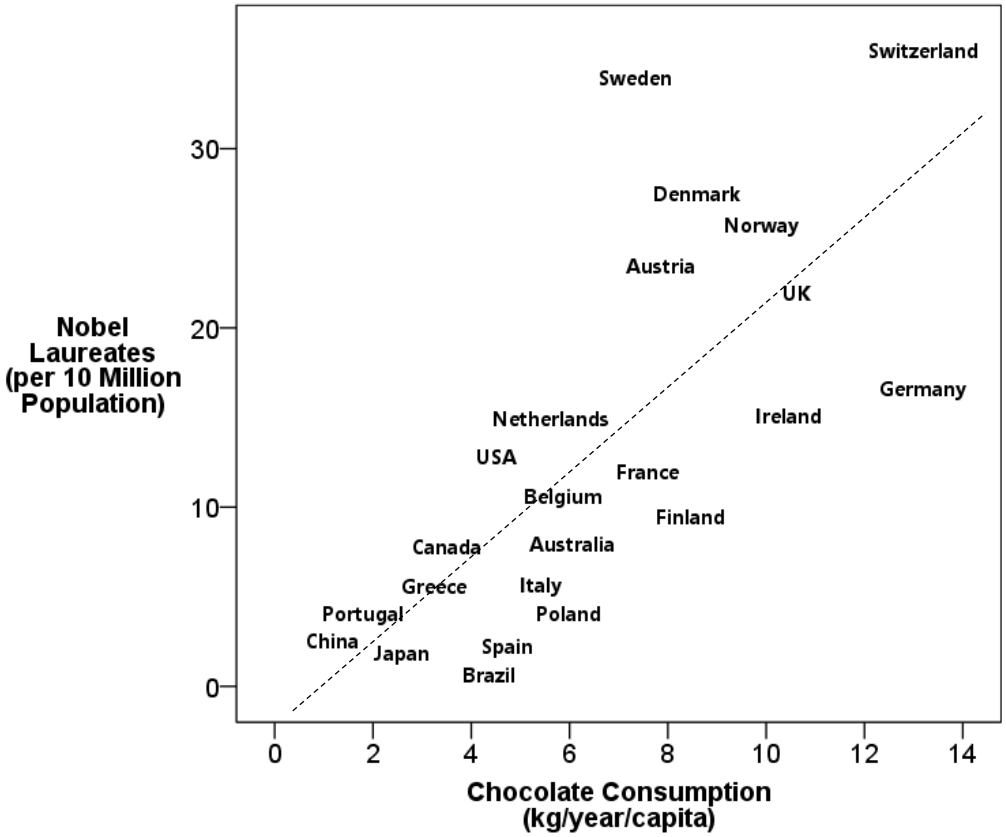
Correlation and Causation:
A Case Study, with Chocolate

The number of Nobel prizes won by a country
correlates strongly with its chocolate consumption.

Cardiologist Franz Messerli created a very effective article for the *New England Journal of Medicine* by connecting chocolate consumption with cognitive ability. On the one hand, his creation beautifully illustrates the need to beware of claims that correlation necessarily means causation. Beyond that, his article offers excellent material for interpreting statistical and graphical findings.

Messerli's Figure 1 is recreated below, if less artfully:



Probably the most prominent threat to the validity of his tongue-in-cheek conclusion involves the way his data are *aggregated*. The *unit of analysis* is country rather than individual. The fact that Switzerland is highest per capita on both Chocolate Consumption and Nobel Laureates doesn't tell us anything about individual people. If

one Swiss citizen devours more chocolate than another, does he or she have a better chance of winning the Nobel? We have no data on that.

Further, at the country level, there might be all sorts of factors that affect a nation's position in this scatterplot. Such factors that could be determining that country's level on both of the charted variables. Since "chocolate" and "Nobel" would now be seen as effects rather than causes, any causal link between them would seem less likely.

Suppose that for this topic we took country-level statistics seriously. The overall correlation coefficient *r* between the two variables is 0.77. But since we have the benefit of the scatterplot, we can see much more. At left, when Chocolate Consumption is below 6 kg/year, *r* is quite strong at about 0.8. For the right-hand portion of the plot, *r* is 0.2. It wouldn't be very accurate to cite only the overall result when the nature of the relationship is so sharply divergent for different subsets of countries.

Let's think further about the nature of the sample. It includes only those 23 countries that had any Nobel winners through 2012. Thus *r* may be distorted by *restriction of range*. It's useful to imagine what the plot would look like if all 200-odd countries were included. Most of them, with a Nobel figure of 0, would be concentrated at the very bottom of the chart. In fact, with mostly zeroes, the Nobel variable would have a drastically *non-normal* or *zero-inflated* distribution, making it less amenable to analysis via correlation at all. This might call for a *nonparametric* method to assess any relationship with another variable.

More on the sample. Geographically, how representative of the world are these 23 countries? Suppose we plotted their positions on a map:

See the problem? A poorly representative sample means poor *external validity*. Findings from Messerli's facetious analysis would hardly be expected to translate to the world outside Europe.

Please enjoy your next chocolate treat, and we all look forward to the data you collect on your cognitive performance.

- *Roland B. Stark, Integrative Statistics, May 2022*

Source: Messerli, F. H. (2012).
Chocolate Consumption, Cognitive Function, and Nobel Laureates.
New England Journal of Medicine 367:1562-1564, DOI: 10.1056/NEJMon1211064.