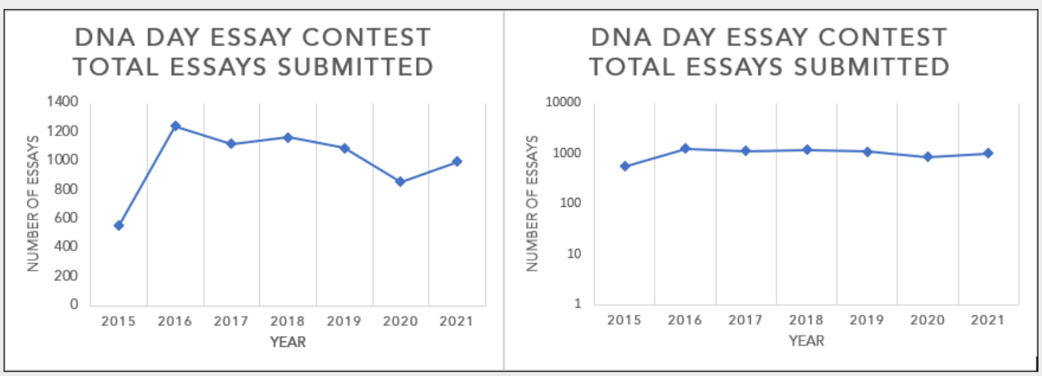
Interpreting Scientific Figures and Statistics

American Society of Human Genetics

Understanding and interpreting scientific figures is a critical skill for scientists. Putting data into a visual format helps scientists analyze, understand, and share their results; however, data visualizations, or the graphic representation of data, can also sometimes be confusing or misleading. With all the data shared on the news and social media which illustrates important issues like disease outbreaks and genetic risks, the ability to critically interpret these data visualizations has become more important than ever. If you look out for a few common errors, you can be a savvier consumer of data visualizations.

Common Error 1: Misleading Axes

Always check the axes of graphs because different axes formats can greatly affect how the data in a graph look. The horizontal or X-axis reads from left to right and the vertical or Y-axis reads bottom to top. For both axes, the scale, or interval between ticks, (e.g., 0-1500 on the left Y-axis below) and the labeled increments (e.g., 0, 500, 1000, 1500 on the left Y-axis below) should be suitable for the data. The graphs below demonstrate how changing the axis scale and increments can make a big difference. Both graphs show the number of essays submitted to the ASHG DNA Day Essay Contest from 2015 to 2021; however, different scales and increments are used. In the left graph, a linear scale is used, and each increment represents 500 essays. In the right graph, a logarithmic scale is used where each y-tick represents an increasing factor of 10. While logarithmic scales can be useful for visualizing data with a very large spread, it is misleading in this case because it makes it look like the number of submissions stayed about the same each year while the linear graph on the left shows that the number of submissions more than doubled from 2015 to 2016.



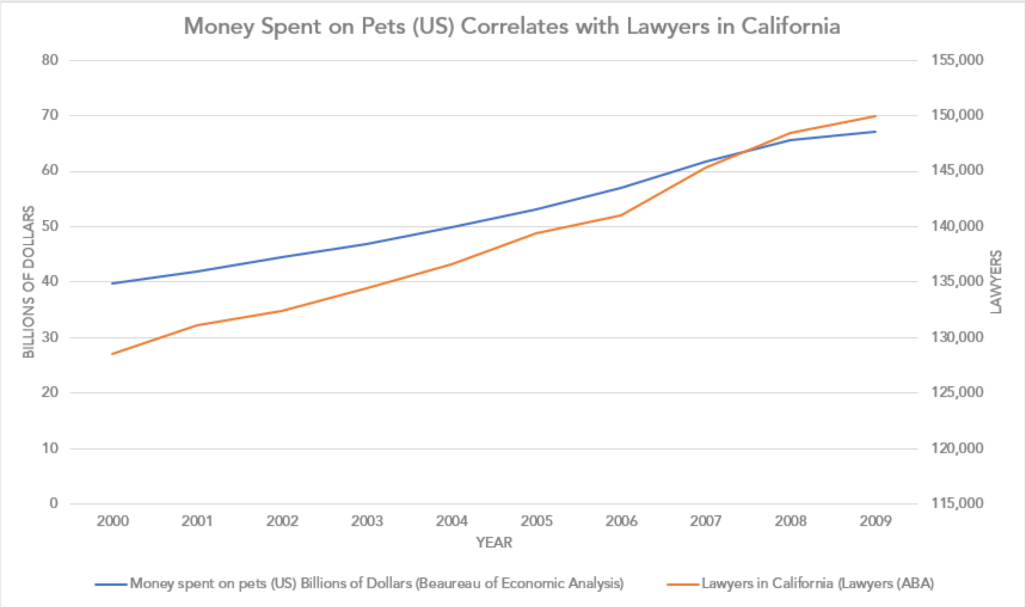
Source: ASHG

Common Error 2: Confusing Correlation with Causation

When two variables are correlated, it means that there is a pattern in the data. So, as one variable changes, the other variable also changes. This may make it seem like the change in one variable is causing the change in the other, but that is not always the case. A strong correlation could indicate causality, but there can be other explanations. In some cases, changes in a third variable could be causing similar changes in the variables shown in the graph. For example, as the temperature increases, the amount of ice cream sold and the number of people with sunburns at the beach will both increase. If you graphed ice cream sales and sunburn on the same graph, some might assume that the increase in one is causing the increase in the other; however, it is really the increase in temperature that is causing the change in the other two variables.

A correlation between two variables could also be due to random chance, where the variables appear to be related but there is no true underlying relationship between them. The graph below plots the amount of money spent on pets (green data) against the number of lawyers in California (blue data) between 2000 and 2009. Both appear to have increased at the same rate over time; however, it is unlikely that either increase caused the other increase.

Money spent on pets vs. the number of lawyers in California from 2000 to 2009



Source: Spurious Correlations

More Questions to Ask

- **Do the visuals (e.g., graph, pictogram) match the numbers?** For example, do the sizes of pieces in a pie graph match the percentages given?
- **When and how was the data collected?** For example, consider data collected about what type of music Americans listen to. You can imagine that the data collected on a college campus may be different from data collected outside an opera hall. Make sure that when and how the data was collected will answer the question being asked.
- **Who funded the study?** Who else is involved in getting the message out? Someone with a vested interest in the data could manipulate the data presentation.