swirl lesson 1: Principles of analytic graphs

Raphael Carvalho

23/03/2020

In this lesson, we'll discuss some basic principles of presenting data effectively. These will illustrate some fundamental concepts of displaying results in order to make them more meaningful and convincing. These principles are cribbed from Edward Tufte's great 2006 book, Beautiful Evidence. You can read more about them at the www.edwardtufte.com website.

As a warm-up, which of the following would NOT be a good use of analytic graphing?

[x] To decide which horse to bet on at the track

[ ] To show causality, mechanism, explanation

[ ] To show comparisons

[ ] To show multivariate data

You're ready to start. Graphs give us a visual form of data, and the first principle of analytic graphs is to show some comparison. You'll hear more about this when you study statistical inference (another great course BTW), but evidence for a hypothesis is always relative to another competing or alternative hypothesis.

When presented with a claim that something is good, you should always ask "Compared to What?" This is why in commercials you often hear the phrase "other leading brands". An implicit comparison, right?

Consider this boxplot which shows the relationship between the use of an air cleaner and the number of symptom-free days of asthmatic children. (The top and bottom lines of the box indicate the 25% and 75% quartiles of the data, and the horizontal line in the box shows the 50%.) Since the box is above 0, the number of symptom-free days for children with asthma is bigger using the air cleaner. This is good, right?

How many days of improvement does the median correspond to?

[ ] -2

[x] 1

[ ] 12

[ ] 4

While it's somewhat informative, it's also somewhat cryptic, since the y-axis is claiming to show a change in number of symptom-free days. Wouldn't it be better to show a comparison? Like this? Here's a graphic which shows two boxplots, the one on the left showing the results for a control group that doesn't use an air cleaner alongside the previously shown boxplot.

By showing the two boxplots side by side, you can clearly see that using the air cleaner increases the number of symptom-free days for most asthmatic children. The plot on the right (using the air cleaner) is generally higher than the one on the left (the control group).

What does this graph NOT show you?

[ ] Children in the control group had at most 3 symptom-free days

[ ] Half of the children in the control group had no improvement

[x] Using the air cleaner makes asthmatic children sicker

[ ] 75% of the children using the air cleaner had at most 3 symptom-free days3

So the first principle was to show a comparison. The second principle is to show causality or a mechanism of how your theory of the data works. This explanation or systematic structure shows your causal framework for thinking about the question you're trying to answer.

Consider this plot which shows the dual boxplot we just showed, but next to it we have a corresponding plot of changes in measures of particulate matter. This picture tries to explain how the air cleaner increases the number of symptom-free days for asthmatic children. What mechanism does the graph imply?

[ ] That the children in the control group are healthier

[ ] That the air cleaner increases pollution

[x] That the air cleaner reduces pollution

[ ] That the air in the control group is cleaner than the air in the other group

By showing the two sets of boxplots side by side you're explaining your theory of why the air cleaner increases the number of symptom-free days. Onward!

So the first principle was to show some comparison, the second was to show a mechanism, so what will the third principle say to show? - Multivariate data!

What is multivariate data you might ask? In technical (scientific) literature this term means more than 2 variables. Two-variable plots are what you saw in high school algebra. Remember those x,y plots when you were learning about slopes and intercepts and equations of lines? They're valuable, but usually questions are more complicated and require more variables. Sometimes, if you restrict yourself to two variables you'll be misled and draw an incorrect conclusion.

Consider this plot which shows the relationship between air pollution (x-axis) and mortality rates among the elderly (y-axis). The blue regression line shows a surprising result. (You'll learn about regression lines when you take the fabulous Regression Models course.). What does the blue regression line indicate?

[ ] As pollution increases the number of deaths doesn't change

[ ] Pollution doesn't really increase, it just gets reported more

[x] As pollution increases, fewer people die

[ ] As pollution increases more people die

Fewer deaths with more pollution? That's a surprise! Something's gotta be wrong, right? In fact, this is an example of Simpson's paradox, or the Yule–Simpson effect. Wikipedia (http://en.wikipedia.org/wiki/Simpson%27s_paradox) tells us that this "is a paradox in probability and statistics, in which a trend that appears in different groups of data disappears when these groups are combined.". Suppose we divided this mortality/pollution data into the four seasons. Would we see different trends?

Yes, we do! Plotting the same data for the 4 seasons individually we see a different result.

What does the new plot indicate?

[ ] As pollution increases fewer people die in all seasons

[ ] Pollution doesn't really increase, it just gets reported more

[ ] As pollution increases the seasons change

[x] As pollution increases more people die in all seasons

The fourth principle of analytic graphing involves integrating evidence. This means not limiting yourself to one form of expression. You can use words, numbers, images as well as diagrams. Graphics should make use of many modes of data presentation. Remembere, "Don't let the tool drive the analysis!"

To show you what we mean, here's an example of a figure taken from a paper published in the Journal of the AMA. It shows the relationship between pollution and hospitalization of people with heart disease. As you can see, it's a lot different from our previous plots. The solid circles in the center portion indicate point

estimates of percentage changes in hospitalization rates for different levels of pollution. The lines through the circles indicate confidence intervals associated with these estimates. (You'll learn more about confidence intervals in another great course, the one on statistical inference.).

Note that on the right side of the figure is another column of numbers, one for each of the point estimates given. This column shows posterior probabilities that relative risk is greater than 0. This, in effect, is a measure of the strength of the evidence showing the correlation between pollution and hospitalization. The point here is that all of this information is located in one picture so that the reader can see the strength of not only the correlations but the evidence as well.

The fifth principle of graphing involves describing and documenting the evidence with sources and appropriate labels and scales. Credibility is important so the data graphics should tell a complete story. Also, using R, you want to preserve any code you use to generate your data and graphics so that the research can be replicated if necessary. This allows for easy verification or finding bugs in your analysis.

The sixth and final principle of analytic graphing is maybe the most important. Content is king! If you don't have something interesting to report, your graphs won't save you. Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.

Review time!!!

Which of the following is NOT a good principle of graphing?

[ ] To describe and document evidence

[ ] To integrate multiple modes of evidence

[x] Having unreadable labels

[ ] Content is king

Which of the following is NOT a good principle of graphing?

[ ] To demonstrate a causative mechanism underlying a correlation [ ] Content is king [ ] To show two competing hypothesis *[ ] To prove you're always right*

Which of the following is NOT a good principle of graphing?

[ ] To show good labels and scales

[ ] Content is king

[x] To show that some fonts are better than others

[ ] To integrate different types of evidence

True or False? Color is king.

[ ] True

[x] False