Trustworthy Data Analysis

Roger Peng 2018/06/04

The success of a data analysis depends critically on the audience. But why? A lot has to do with whether the audience *trusts* the analysis as well as the person presenting the analysis. Almost al presentations are incomplete because for any analysis of reasonable size, some details must be omitted for the sake of clarity. A good presentation will have a structured narrative that will guide the presenter in choosing what should be included and what should be omitted. However, audiences will vary in their acceptance of that narrative and will often want to know if other details exist.

The Presentation

Consider the following scenario:

> A person is analyzing some data and is trying to determine if two features, call them X and Y, are related to each other. After looking at the data for some time, they come to you and declare that the Pearson correlation between X and Y is 0.85 and therefore conclude that X and Y are related.

The question then is, do you trust this analysis?

Given the painfully brief presentation of the scenario, I would imagine that most people experienced in data analysis would say something along the lines of "No", or at least "Not yet". So, why would we not trust this analysis?

There are many questions that one might ask before one were to place any trust in the results of this analysis. Here are just a few:

- What are X and Y? Is there a reason why X and Y might be causally related to each other? If mere correlation is of interest, that's fine but some more detail could be illuminating.
- What is the sampling frame here? Where did the data come from? We need to be able to understand the nature of uncertainty.
- What is the uncertainty around the correlation estimate? Are we are looking at noise here or a real signal. If there is no uncertainty (because there is no sampling) then that should be made clear.
- How were the data processed to get to the point where we can compute Pearson's correlation? Did missing data have to be removed? Were there transformations done to the data?
- The Pearson correlation is really only interpretable if the data X and Y are Normal-ish distributed. How do the data look? Is the interpretation of correlation here reasonable?
- Pearson correlation can be driven by outliers in X or Y. Even if the data are mostly Normal-ish distributed, individual outliers could make the appearance of a strong relationship even if the bulk of the data are not correlated. Were there any outliers in the data (either X or Y)?

The above questions about the presentation and statistical methodology are all reasonable and would likely come up in this scenario. In fact, there would likely be even more questions asked before a one could be assured that the analysis was trustworthy, but this is just a smattering.

Done but Not Presented

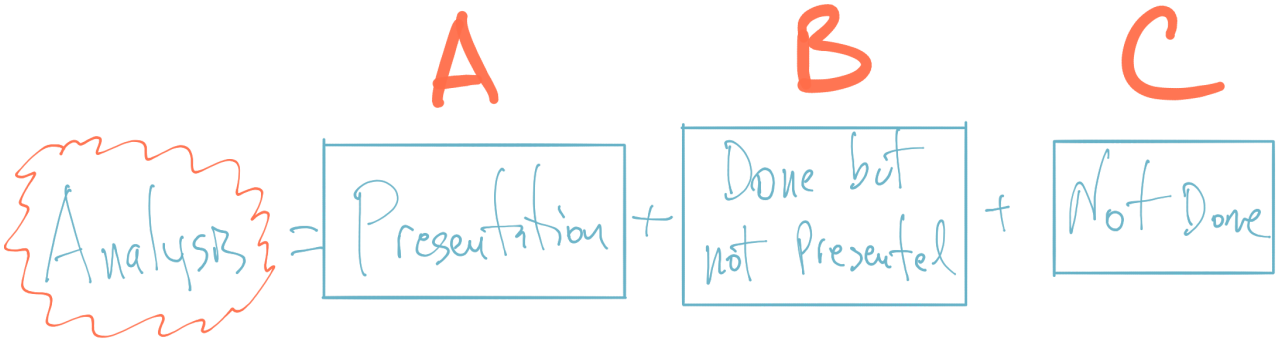
I think it's reasonable to think that a *good* analyst would have concrete answers to all of these questions even though they were omitted from the presentation.

- They would know what X and Y were and whether it made sense to determine if they were correlated, based on the context of the data and questions being asked.
- They would know how the data came to them, whether they represented a sample, and what the population was.
- They would have calculated the uncertainty around the correlation estimate (e.g. a 95% confidence interval) and would have it on hand to present to you.
- Ideally, they would have processed the data themselves and would be able to explain what had to be done in order to get the data to the point where correlations could be computed. If they didn't process the data, they would at least be able to describe what was done and whether those actions have a major impact on the results.
- To justify the use of Pearson correlation, they would have made a histogram (or a Q-Q plot) to look at the marginal distributions of X and Y. If the data weren't Normal looking they would have considered some possible transformations if possible.
- To check for outliers, a scatterplot of X and Y would have been made to examine if any small number of points was driving the observed correlation between X and Y. Even though they didn't show you the scatterplot, they might have it on hand for you to examine.

One might think of other things to do, but the items listed above are in direct response to the questions asked before.

Done and Undone

My "analysis of variance" representation of a data analysis is roughly



Here we have

- **A**: The presentation, which in the above example, is the simple correlation coefficient between X and Y
- **B**: The answers to all of the questions that likely would come up after seeing the presentation
- **C**: Anything that was not done by the analyst

We can only observe A and B and need to speculate about C. The times when I most trust an analysis is when I believe that the C component is relatively small, and is essentially orthogonal to the other components of the equation (A and B). In other words, were one to actually do the things in the "Not Done" bucket, they would have no influence on the overall results of the analysis. There should be nothing surprising or unexpected in the C component.

No matter what data is being analyzed, and no matter who is doing the analysis, the **presentation** of an analysis must be limited, usually because of time. Choices must be made to present a selection of what was actually done, therefore leaving a large number of items in the "Done but not Presented" component. An analogy might be when one writes slides for a presentation, often there are a few slides that are left in the back of the slide deck that are not presented but are easily retrieved should a question come up. The material in those slides was important enough to warrant making a slide, but not important enough to make it into the presentation. In any substantial data analysis, the number of "slides" presented as the results is relatively small while the number of "slides" held in reserve is potentially huge.

Another large part of a data analysis concerns *who* is presenting. This person may or may not have a track record of producing good analyses and the background of the presenter may or may not be known to the audience. My response to the presentation of an analysis tends to differ based on who is presenting and my confidence in their ability to execute a good analysis. Ultimately, I think my approach to reviewing an analysis comes down to this:

1. If it's a first presentation, then regardless of the presenter, I'll likely want to see A and B, and discuss C. For a first presentation, there will likely be a number of things "Not Done" and so the presenter will need to go back and do more things.
2. If we're on the second or third iteration and the presenter is someone I trust and have confidence in, then seeing A and part of B may be sufficient and we will likely focus just on the contents in A. In part, this requires my trust in their judgment in deciding what are the relevant aspects to present.
3. For presenters that I trust, my assumption is that there are many things in B that are potentially relevant, but I assume that they have done them and have incorporated their effects into A. For example, if there are outliers in the data, but they do not seem to introduce any sensitivity into the results, then my assumption is that they looked at it, saw that it didn't really make a difference, and moved on. Given this, my confidence that the elements of C are orthogonal to the results presented is high.
4. For presenters that I'm not familiar with or with whom I have not yet built up any trust, my assumptions about what lies in B and C are not clear. I'll want to see more of what is in B and my skepticism about C being orthogonal to the results will be high.

One of the implications of this process is that two different presenters could make the *exact same presentation* and my response to them will be different. This is perhaps an unfortunate reality and opens the door to introducing all kinds of inappropriate biases. However, my understanding of the presenters' abilities will affect how much I ask about B and C.

At the end of the day, I think an analysis is *trustworthy* when my understanding of A and B is such that I have reasonable confidence that C is orthogonal. In other words, there's little else that can be done with the data that will have a meaningful impact on the results.

Implications for Analysts

As an analyst it might be useful to think of what are the things that will fall into components A, B, and C. In particular, how one thinks about the three components will likely depend on the audience to which the presentation is being made. In fact, the "presentation" may range from sending a simple email, to delivering a class lecture, or a keynote talk. The manner in which you present the results of an analysis is *part of the analysis* and will play a large role in determining the *success* of the analysis. If you are unfamiliar with the audience, or believe they are unfamiliar with you, you may need to place more elements in components A (the presentation), and perhaps talk a little faster. But if you already have a long-term relationship with the audience, a quick summary (with lots of things placed into component B) may be enough.

One of the ways in which you can divide up the things that go into A, B, and C is to develop a good understanding of the audience. If the audience enjoys looking at scatterplots and making inquiries about individual data points, then you're going to make sure you have that kind of detailed understanding in the data, and you may want to just put that kind of information up front in part A. If the audience likes to have a higher level perspective of things, you can reserve the information for part B.

Considering the audience is useful because it can often drive you to do analyses that perhaps you hadn't thought to do at first. For example, if your boss always wants to see a sensitivity analysis, then it might be wise to do that and put the results in part B, even if you don't think it's critically necessary or if it's tedious to present. On occasion, you might find that the sensitivity analyses in fact sheds light on an unforeseen aspect of the data. It would be nice if there were a "global list of things to do in every analysis", but there isn't and even if there were it would likely be too long to complete for any specific analysis. So one way to optimize your approach is to consider the audience and what they might want to see, and to merge that with what you think is needed for the analysis.

If you *are* the audience, then considering the audience's needs is a relatively simple task. But often the audience will be separate (thesis committee, journal reviewers/editors, conference attendees) and you will have to make your best effort at guessing. If you have direct access to the audience, then a simpler approach would be to just ask them. But this is a potentially time-consuming task (depending on how long it takes for them to respond) and may not be feasible in the time frame allowed for the analysis.

Trusting vs. Believing

It's entirely possible to trust an analysis but not believe the final conclusions. In particular, if this is the first analysis of it's kind that you are seeing, there's almost no reason to believe that the conclusions are true until you've seen other independent analysis. An initial analysis may only have limited preliminary data and you may need to make a decision to invest in collecting more data. Until then, there may be no way to know if the analysis is *true* or not. But the analysis may still be trustworthy in the sense that everything that should have been done was done.

Looking back at the original "presentation" given at the top, one might ask "So, is X correlated with Y?". Maybe, and there seems to be evidence that it is. However, whether I ultimately believe the result will depend on factors outside the analysis.

You can hear more from me and the JHU Data Science Lab by subscribing to our weekly newsletter Monday Morning Data Science.

‹
Context Compatibility in Data Analysis
Estimating mortality rates in Puerto Rico after hurricane María using newly released official death counts ›
