Visualizing Uncertainty due to Missing Data

Hanne Ida Oberman*

Utrecht University

1 Introduction

Familiarity with visualized uncertainty is one of the few unexpected positive effects of the COVID-19 crisis on society at large. The average layperson was never before exposed to so many scientific data visualizations. Whether it is in daily updates of COVID prognoses, or in comparing models for potential restrictions to implement–each is expressed with a proper range of ‘unknowns’. Media consumers might have seen ranges in predictions before, such as in exit polls of electoral votes, but never was uncertainty so omnipresent. This societal need for *and* interest in visualized uncertainty is unprecedented.

Now that expressing uncertainty has breached the gap between visualizations made for analysis purposes and visualizations for communication, we should keep the ball rolling. The obscurity about future developments that is apparent in (COVID) prediction models is not the only possible source of uncertainty one could plot. Take for example confidence intervals, with which scientists express uncertainty about their hypotheses when extrapolating from a given sample to the greater population. These are often represented by error ranges in study results. In a study investigating the efficacy of two different treatment options, one may plot a bar graph with confidence intervals as error bars on the treatment effects. This is an intuitive way of showing the audience something about the strength of the evidence presented–much more so than reporting confidence intervals and other statistics textually. And unsurprisingly, textual expressions of uncertainty rarely make the headlines. With that, the magnitude of the uncertainty stays exclusively known to the scientific community. Which, in turn, may cause the wider public to distrust scientific studies. This gives us all the more reason to express uncertainty visually.

But, even if there is no uncertainty due to study design, such as in de US census, there may still be unknown factors that should be expressed visually. A ubiquitous and often ignored problem in data analyses is missing data. Missingness can lead to severely biased results and increases the room for erroneous conclusions [1]. Missing data may occur across observations (e.g., some people have no home address and cannot be reached) or within observations (e.g., some topics are deemed too sensitive and will not elicit complete responses). Whatever the reason of the missingness, it may gravely influence any subsequent estimates [2]. Uncertainty due to missingness should therefore be expressed in any visualization for communication.

2 Methods

In this project, we focus on intuitive ways to express uncertainty due to missing data to non-expert audiences. There have been many methodological advancements in the area of missing data, but none have specifically focused on visualization. We base our work on the method that has become the ‘gold standard’ for mitigating missingness: ‘imputing’ (i.e., filling in) any missing entries before analyzing incomplete datasets [3][4][5]. How this method works exactly is not the focus of the current project. Instead, we aim to develop an online evaluation suite to inspect missing data, impute it, and evaluate it. At each of these three steps, there should be data visualization tools that are understandable to novice and lay audiences. This is where we would love your help: what do *you* think are intuitive ways to express uncertainty due to missing data?

3 Questions for Attendees

Could you please provide feedback on our tools for visualizing missing data? We have developed a pilot version of our missing data evaluation app, which is available online. We do not assume that you, attendee to this conference, are in any way familiar with the methodology in this app. The intended audience of the app are applied researchers who are faced with missing data. We want to enable them to easily produce understandable graphical presentations of the uncertainty in their data. What we would like to know from you is whether these visualizations are *indeed* understandable. At some point, you might encounter these figures in visualizations for communication. But before that, we would like feedback from an audience that is not an expert in methodology, but rather in data visualization. What aspects are unclear to you? What could we do to improve our missing data visualization app?

Please navigate to the online pilot app and look around. As a motivating example, the pilot includes incomplete data on child development. The most instructive variables to use, arguably, are height (`hgt`) and weight (`wgt`). These variables have an intuitive, apparent relationship, and do not share complex overlapping missing values. Feel free to click around and enjoy! For the interested reader, there is more information available at stefvanbuuren.name/fimd [5].

References

[1] Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (2nd ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119013563

[2] Graham, J. W. (2012). Missing data: Analysis and design. Springer Science & Business Media.

[3] Rubin, D. B. (1976). Inference and Missing Data. Biometrika, 63(3), 581–592. JSTOR. https://doi.org/10.2307/2335739

[4] Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147.

[5] Van Buuren, S. (2018). Flexible imputation of missing data. Chapman and Hall/CRC. https://stefvanbuuren.name/fimd/

* e-mail: h.i.oberman@uu.nl