Version control as a learning objective in statistics and data science courses

Maria Tackett[1]; Matthew Beckman[2]; Mine Çetinkaya-Rundel[1]; Nicholas Horton[3]; Colin Rundel[1]; Adam Sullivan[4]

1   Duke University, Durham, NC, USA
2   Pennsylvania State University, State College, PA, USA
3   Amherst College, Amherst, MA, USA
4   Takeda, Cambridge, MA, USA

Abstract:

The *2014 American Statistical Association Curriculum Guidelines for Undergraduate Programs* included the development of computing skills specific to well-documented and reproducible analyses as a necessary component of the undergraduate statistics curriculum. Given data analysis is an iterative process that often relies on and produces many files (Bryan, 2018), version control is foundational for reproducible workflows. Including version control as a learning objective in introductory statistics and data science courses helps equip students with the computing skills required to conduct reproducible analyses, both individually and collaboratively. Additionally, as version control systems, such as Git, are widely used in academia and industry, it is an opportunity for students to start developing the skills they'll use in research, internships, and their future careers.

In this talk, I will describe how version control with Git can be included as a learning objective in the first and second statistics course. I'll discuss strategies for introducing version control to students, incorporating it in individual and team-based assignments, and assessing students' understanding. I'll also share lessons learned and an example of how this can be implemented using RStudio and GitHub.