STAT 542: Computing for Data Science

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2025

Chapter 1: Introduction to Data Science

- Our book defines **data science** as the science of extracting meaningful information from data.
- Data science — a combinations of concepts:
  1. Combines computer science (abstract computational structure and development of algorithms) with statistics (sampling, models, distributions, decision-making).
  2. Combines mathematical/statistical/computational skill with knowledge of the domain the data came from.
  3. Distinction between data (abundant) and useful information (may be scarce).

Relationship of Traditional Statistics to Data Science

- Traditional statistical methods were developed as a way to obtain information from scarce or limited data.
- In many settings, data is now plentiful: Need modern data science methods.
- Modern data sets may be *observational* and may not come from *random samples*.
- So traditional probability models may not work for such data.

Appendix B: A brief introduction to R

- Appendix B introduces some very basic aspects of R programming.
- Read this and then do the first homework assignment!
- Key topics:
  1. Different object types in R: Vectors, lists, matrices, dataframes, tibbles
  2. Attributes and classes of objects
  3. Element types in vectors: logical (TRUE/FALSE), character (strings), integer (whole numbers), double (real numbers). Less common: complex, raw
  4. Factors and how they differ from character vectors
  5. Using functions and their arguments
  6. Using add-on packages
- This material will be review for many of you.