Course Introduction

"…it is a function of statistical method to emphasize that precise conclusions cannot be drawn from inadequate data."

- E.S. Pearson and H.O Hartley

New! Fall 2020
Dear BIOSTATS 690C Fall 2020,

This year, BIOSTATS 690C, in ***addition to Stata***, will also an introduction to R using ***RStudio***.
Please be aware that this is a "work in progress" as I transition this course from Stata- to R-based.
Thank you – Carol

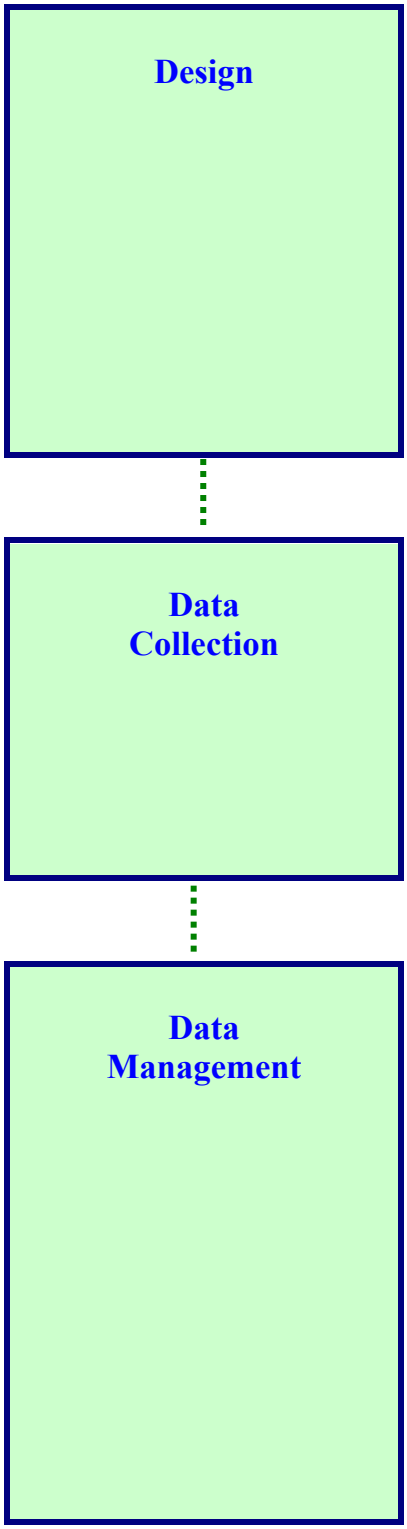
This course introduction is a road map of the entire course. It places *data management and statistical computing* in the larger context of scientific research and publication. A brief overview of each unit is also provided.

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

Table of Contents

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

1. Course Roadmap



It is unethical to do bad science. Valuable resources (including human subjects) would be misused and incorrect conclusions can have bad consequences. Study ***design*** encompasses all the structural features of research: the research question, design type (e.g. – randomized, case-control, cross-sectional, etc.), sampling/selection of study sample, sample size, and choice of study variables (predictor, outcome, covariate).

This course is ***not*** about design.

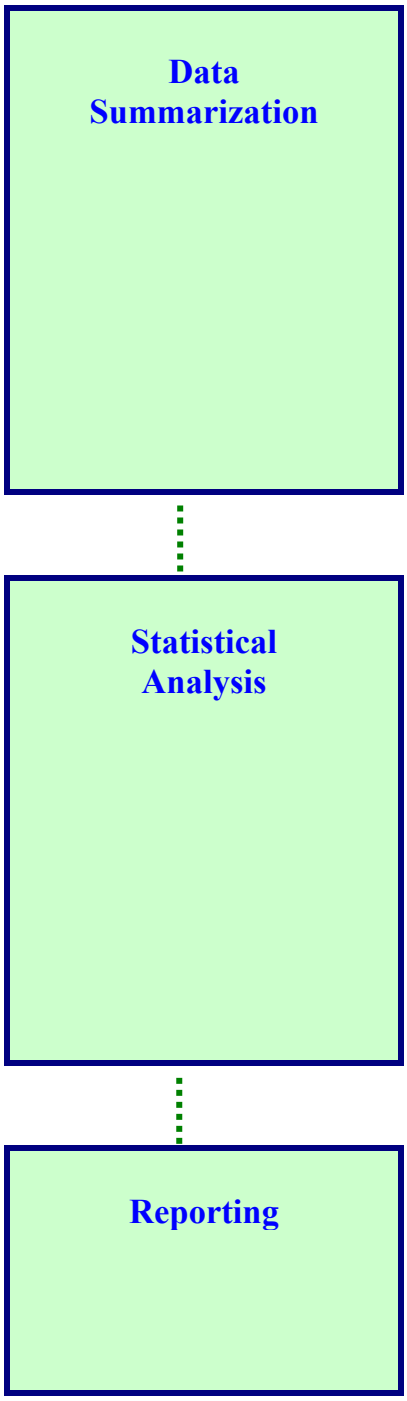
As discussed in BIOSTATS 540, *Introductory Biostatistics*, recorded data represent a selection of what has been observed. This course is ***not*** about the choice of data to collect nor the scheme for assigning data values.

Rather, a focus of this course is methods for the ***quality of data collection***.

Data management encompasses all aspects of data processing and serves multiple purposes – protection of human subject confidentiality, cleaning and validation, transformation and readying for statistical analysis, and backup.

This course emphasizes ***good practices for data collection and management***.

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting



The goal of ***data summarization*** is to communicate the important aspects of the data, ***simply and as clearly as possible***. Summaries of location (mean, median, etc.) and dispersion (variance, standard deviation) are perhaps familiar. Often important features of data are its distributional shape and the patterns (if any) of outlying values.

This course introduces the use of Stata (version 16, but 14 or 15 are fine) for obtaining both numerical and graphical summaries of data. ***Graphical summaries are emphasized***.

The goal of ***statistical analysis*** is to discover relationships (e.g. – weight increases with height) and to compare alternative explanations for the data (e.g. – a treatment does or does not work). It involves the fitting of models to the data (not the other way around!) and is followed by their assessment.

This course provides an extension of BIOSTATS 640, *Intermediate Biostatistics*. It introduces the use of Stata version 16 for one and two sample hypothesis tests, normal theory linear regression, categorical data analysis, and logistic regression.

Text, tables and graphs are the three tools used for reporting. This course is ***not*** about the preparation of publications.

However, this course does provide suggestions for the design of tables and graphs.

2. Overview, Unit by Unit

Unit 1 - Principles of Data Management

Unit 1 is an introduction to several aspects of data management:

- Data management plan
- Data Collection Instrument
- Data Entry and documentation
- Variable creation, verification, and documentation
- Database structure (eg flat versus relational)
- Data Archiving

Unit 2 - Ethical Management of Human Subjects Data

Unit 2 emphasizes the protection of human subject confidentiality. You will learn about the Health Insurance Portability and Accountability Act (HIPAA). You will also learn strategies for de-identifying human data records. As part of this unit, you will be required to complete the "Basic Course" offered by the Collaborative Institutional Training Initiative (CITI). Successful completion (CITI certification) is required by most research institutions.

Unit 3 - MS Excel for Epidemiology

While MS Excel is unlikely to be your "workhorse" for data management (fingers crossed, I can equip you with a bigger arsenal of tools), it's certainly useful to be skilled in MS Excel! This introduction to MS Excel focuses on its use for data set creation, manipulation (e.g. – sorting and selecting) and summarization (e.g. – mathematical calculations such as sums, differences, and functions). The use of MS Excel for a small number of useful graphical summaries is also described. Just as MS Excel is unlikely to be your "go to" for data management, it is also ***not emphasized*** for data analysis either. We will be learning Stata (or maybe some R/RStudio) for data analysis.

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

Unit 4 -
Introduction to Stata (or R using RStudio)

This unit is an introduction to the basics of Stata version 16 (hereafter referred to as Stata), or R/RStudio, and its use in data management:

- Launching and exiting Stata (or R/RStudio)
- Using Stata as a calculator (mathematics, probability calculations, etc)
- Basic commands for working with observations (sorting, selecting, deleting)
- Basic commands for working with variables (creating, documenting)
- Working with a single data set – input and output
- Working with multiple data sets – concatenation, merging

Unit 5 -
Stata (or R using RStudio) for Data Description

The use of Stata for numerical summaries is discussed. You will learn the syntax for obtaining standard descriptive statistics, including – means, medians, variance, standard deviations, and percentiles. You will also learn how to use Stata to obtain confidence intervals, etc.

Unit 6 -
Stata (or R using RStudio) for Graphs

The use of Stata for producing graphical summaries of data is discussed and their importance is emphasized. You will learn how to produce several graphs including, but not limited to: bar graphs, histograms, X-Y scatterplots, repeated measures profiles, and summary repeated measures profiles.

Unit 7 -
Stata (or R using RStudio) for Analysis of 1, 2 and 3^+ Samples

The use of Stata for confidence interval estimation and hypothesis tests in the one, two and more than two sample setting is introduced. This discussion includes both parametric (z-test, t-test, F-test for analysis of variance) and non-parametric procedures (Signed Rank, Wilcoxon Rank Sum).

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

Unit 8 -
Stata (or R using RStudio) for Categorical Data Analysis

In this unit, you will be introduced to the use of Stata for standard analyses of epidemiologic data in the form of counts. You will learn the syntax for the analysis of 2x2 tables, RxC tables, and stratified 2x2 tables. This unit includes a review of the interpretation of computer output and the nature of the conclusions that can be drawn.

Unit 9 -
Stata (or R using RStudio) for Normal Theory Regression

In this unit, you will be introduced to the use of Stata for the fitting and assessment of single and multiple predictor normal theory regression models. In this setting, the outcome variable is continuous and assumed to be normally distributed. You will learn the syntax for: (i) assessing the reasonableness of underlying assumptions; (ii) model fitting and (iii) assessing model adequacy (both numerical and graphical). This unit also includes a review of the interpretation of computer output and the nature of the conclusions that can be drawn.

Unit 10 -
Stata (or R using RStudio) for Logistic Regression

In this unit, you will be introduced to the use of Stata for the fitting and assessment of single and multiple predictor logistic regression models. In this setting, the outcome variable is binary discrete (yes/no) and is assumed to be distributed Bernoulli. Or the data are assumed to be distributed Binomial. You will learn the syntax for: (i) assessing linearity on the logit scale; (ii) model fitting and (iii) assessing model adequacy (both numerical and graphical). You will learn the syntax for model fitting and procedures (both numerical and graphical) for assessing goodness of fit. This unit also includes a review of the interpretation of computer output and the nature of the conclusions that can be drawn.