Introduction to EDF 9770 (and to Quantitative Methods)

- Topics:
  - Why might you be here?
  - The truth about “statistics”
  - Course requirements, responsibilities, and your experience
  - About the R software used in this course
  - What you are supposed to know already (or should review)
  - What we will cover this semester (and what could be next)

Two Reasons Why You Might Be Here

- "**This class is required**" (and I just need to pass it).
  - I get it—and it's ok if this is the only reason you are here, but I hope to convince you otherwise!

- "**I want to learn more about analysis of quantitative data**"!
  - One method by which to answer questions—in research settings or in real life—is by collecting quantitative data
  - The process of summarizing that data—by finding patterns in order to answer questions—requires statistical models
  - Quantitative methods = Quantitative data + application of statistical models to answer questions

“Statistics” Gets a Bad Rap

- **Statistics = applied math** used for a relevant purpose!
  - Btw, “**data science**” is the more modern label for “statistics” (*but it often emphasizes prediction more than theory testing*)

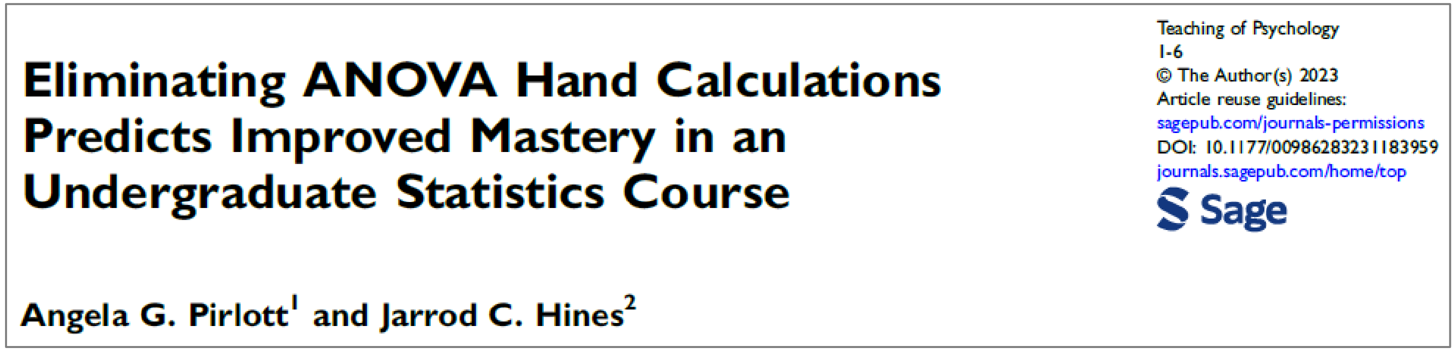
- Competent consumers and users of quantitative methods must learn the **language and logic** behind the uses of statistical models

- This will **NOT** require anxiety-provoking behaviors like:
  - **Deriving formulas or results**—it’s ok to trust the people who specialize in these areas to have gotten it right and use their work (for now, at least)
  - **Memorizing formulas**—it’s ok to trust the computer programmers who have implemented various estimation techniques (for now, at least)
  - **Calculating things by hand!** Because…

Why No Hand Calculations? 4 Reasons

- Manual computations → **error**, and computers are always better and faster

- It doesn’t help you learn as effectively (as using software)



Teaching of Psychology
1-6
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00986283231183959
journals.sagepub.com/home/top
Sage

Eliminating ANOVA Hand Calculations Predicts Improved Mastery in an Undergraduate Statistics Course

Angela G. Pirlott[1] and Jarrod C. Hines[2]

- That’s now how analyses are done for **real-life** purposes (which is what I am choosing to emphasize)

- More advanced analyses **cannot be done by hand** anyway (i.e., they require iterative estimation methods)

The Truth about “Statistics”

- The hardest part about learning statistics is not the math… it is the working memory load of new language + logic!

- **Language**: Ideas will be expressed through words, notation (symbols in equations), and computer code (“syntax”)

- **Logic**: Decision guidelines for matching data types and questions to statistical models (and then “estimating” models)

- **Working memory load** is reduced through frequent exposure, mindful repetition, and engagement → automaticity
  - Revising this material once a week is likely not enough
  - Our material builds cumulatively, so staying checked in will help!

How I Will Help You Acquire the Language and Logic of Statistical Modeling

- I believe that everyone is capable and can significantly benefit** from learning how to use quantitative methods!

- Philosophy: Focus on accessibility + mastery learning

- **Materials**: Unit = (wordy) lecture + example(s); 5 planned (+ 2 more lectures)
  - **Lecture** slides present concepts—the **what** and the **why**
  - **Example** documents: reinforce the concepts and demonstrate the **how** using **R** software
    - Unit 1 will have an R software demo video instead of an example document
  - All available at the course website (hosted externally to Canvas for your future selves)

- ** *Benefits include but are not limited to:*
  Better research, more authorship opportunities, and actual money

Course Requirements (due Sunday before class)

- **Formative assessments (FAs)** to provide a structured review in next class
  - Concepts, vocabulary, notation, practice with interpretation
  - **7** planned; **2** points each for **completion** + up to **1** point for **accuracy**

- **Homework (HW)** to practice analyzing data and interpreting results
  - Based **directly on examples** given (no googling or AI required)
  - You will each have a unique dataset (with a common story)
  - **Computational** questions: Instant feedback, infinite attempts
  - **Results** questions: Delayed feedback, multiple choice (so single attempt)
  - **5** planned; up to **59** points for completing them **accurately**
    - Complete "HW 0" over the syllabus for 1 point of extra credit

More About the Course Requirements

- **Individual Project:**
  - **Report analyses of real data in APA-style mini-paper**
  - Ideally on **data you care about** (*less ideally on other public-use data*)
  - Predict 1–2 numeric outcome(s) from 2–3 predictor variables
  - Plan outline due **3/29** (so you must find your data by then)
  - Report document due **4/19**; revision due **5/1**
  - Rubric will be provided along with plan feedback

- **Late work** will be accepted through 5/1 with a penalty per activity:
  - **−1** for FA or project plan, **−3** for HW, **−5** for project report
  - Extensions granted if requested **at least 2 weeks** in advance
  - Due dates may be pushed later if needed (but never sooner)

Our Other Responsibilities

- **My job** (besides providing materials and activities) **is to answer questions:**
  - Via email, in individual meetings, or in group-based zoom office hours
  - Work on HW during office hours and get prompt assistance (no appt needed)

- **Your job** (in descending order of timely importance):
  - Frequently **review** the class material, focusing on mastering the vocabulary (words and symbols), logic, and concepts
  - **Don't wait** until the last minute to start homework; **ask for help**
    - Please email a screenshot of your code+error so I can respond easily
  - Do the **readings** for a broader perspective and more examples (best *after* lecture)
  - **Practice** using the software on **data you care about**!

More About Your Experience in this Class

- **Attendance:** Strongly recommended but not required
  - **You choose** (for any reason): In-person “**roomer**” or “**zoomer**”
  - Please do not attend in-person if you are seriously ill!
  - You won’t miss out: I will post YouTube-hosted recordings (audio + screenshare only) for each class at the course website
    - Videos and generated transcript are searchable!
    - **Videos are best used for specific review**, not in place of attendance (too passive)
  - Ask questions aloud or in the zoom chat window (+DM) in either modality

- **Changes** will be announced via email and Canvas by 9 AM on class days
  - I will change to zoom-only if I am sick!
  - I will change to zoom-only for dangerous weather (less likely now ☺)
  - Nothing is more important than our health and safety…

Class-Sponsored Statistical Software: R

- We are using **R software (via RStudio)** for your future learning
  - **Pros:** Free! Install it on any operating system! Object-oriented!
  - **Cons:** You get what you pay for! *(terrible documentation; inconsistency)*
  - We will write **code** (“**syntax**”) for greater efficiency and reproducibility
  - I will make a video showing how to use R (that goes with Lecture 1)

- **Why not SPSS or JMP?** Because they aren’t used in advanced contexts
  - New models or computational shortcuts usually appear in R packages
  - Btw, SPSS is used in the textbook, and it can do most of our content
  - Btw, JASP is a windows-driven version of R that may be more friendly

No Programming Experience Required

- Don’t worry—I DO NOT need you to memorize code, ever!
  - I will also give you “**functions**” (mini-programs) to better format R output and to automate tedious coding for additional calculations

- **To do HW**, you can do exactly what I (still) do:
  - **Find** the corresponding example I gave you of what you need to do
  - **Determine** how to modify that code to work for your HW
  - **Copy** (control+C), paste (control+V), and **find and replace** (control+H) will be your best friends (Mac: swap control for command)

- Please ask me for help! Show or send me screenshots of your code, error(s), and question(s) you are trying to answer

What You Are Supposed To Know Already

- Listed pre-requisite: EDF 9270 (or equivalent)

- **Working pre-requisites** are familiarity (via stats software) with:
  - Descriptive statistics (e.g., frequency, mean, variance)
  - Bivariate associations (e.g., Pearson correlation)
  - Statistical concepts (e.g., null hypothesis testing)
  - We will quickly review these concepts in units 1–2

- Most of this class will focus on the **GLM**… so what's that?
  - Rather than present as specific cases, I will present the same material in a **model-driven** way to facilitate your acquisition of future content

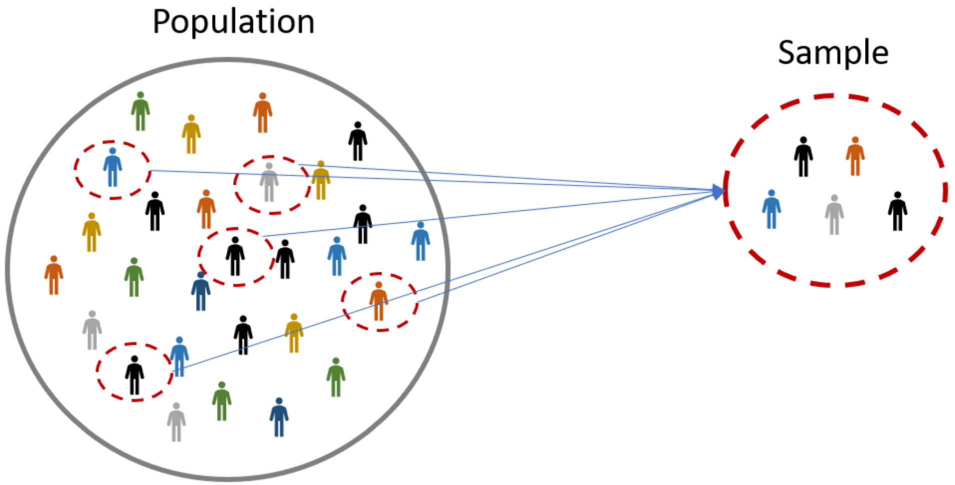
General Linear Models (GLMs) This Semester

- One-stop shop for **predicting one numeric outcome per person** (or “unit”)
  - **Quantitative** predictors = “(linear) regression”
    - 1 numeric predictor variable = “simple (linear) regression” (unit 2)
    - 2+ numeric predictor variables = “multiple (linear) regression” (unit 4)
    - We will cover both linear and nonlinear patterns of relationships (unit 3)
  - **Categorical** predictors = “analysis of variance (ANOVA)”
    - 1 two-group predictor variable = “independent-samples t-test” (unit 2)
    - 1 three-or-more-group predictor variable = “one-way ANOVA” (unit 3)
    - 2+ group predictor variables = “two-way (or factorial) ANOVA” (unit 5)
  - Both kinds of predictors = “analysis of covariance (ANCOVA)”
  - Unit 5 will cover moderation (via interactions) of all kinds, too!
  - Lectures 6 and 7 will set the stage for advanced quantitative methods

So What Kind of Data Can Use GLMs? Let’s Review Some Sampling Vocabulary

- Who are we trying to know about, more generally? → To what population do we want to make inferences?

- Accordingly, from whom should we collect data? → And what info should we collect in our selected sample?
  - Variables are characteristics that differ across units* in a sample



* Units = persons, organizations, animals, etc.

Where to Begin? Sampling Vocabulary

- Example: Let's say a researcher wants to examine graduate student life, so they use a survey to collect self-report info on program membership, stress, and well-being

- So what type of sample should we collect? For instance:
  - Data for multiple students from the same program? Program is a **constant**, not a **variable**
  - To examine **differences between programs**, we'd need to sample **multiple programs** from the same college, at a minimum
  - But would it help our **generalizability** to include multiple colleges from the same university, or even from multiple universities?
  - Should we survey each student **once**? Or would **several times** be better?
  - Should we also try to collect **corresponding data** from other people who know each student well (e.g., their partners, friends, family)?
- These questions address **independent** versus **dependent** sampling…
  - The latter cases are also known as "**dependent data**" for which **GLMs should not be used**!

Independent vs. Dependent Samples

- **The GLM is designed for independent samples!**
  - Example: multiple students in the same program each measured on one occasion
  - If program is a constant, not a variable, it can't be part of any research questions *(but then program differences are controlled as a potential source of variation)*

- Examples of **dependent (= naturally related) samples** (beyond the GLM):
  - Sample **multiple programs** (e.g., >20) from same university
    - e.g., Stress rates of persons from the same program may be more related (dependent) than those of persons from different programs
    - This is known as "**clustered**" or "**nested**" or "**hierarchical**" data
  - Sample each person **more than once** (multiple occasions or conditions)
    - e.g., Stress rates on occasions from the same person may be more related (dependent) than those of occasions from different persons
    - This is known as "**repeated measures**" or "**longitudinal**" data
  - Collect both self-report and other-report ratings → "**dyadic**" data

Wrapping Up

- **End goal of this semester: Learn how to use general linear models** [**GLMs**; *with variants known as regression, analysis of (co)variance*] to analyze quantitative research data
  - Requires learning new logic and language (words, symbols, and code) by which to link data, questions, and models
  - Begin by reviewing how to summarize variables (in Lecture 1) to get to know R software using more familiar ideas
  - Continue with GLMs: statistical models for predicting numeric variables from any kind of variable in independent samples (*which need extensions to be covered elsewhere for predicting other kinds of variables or for use in dependent samples*)

- We will estimate GLMs using **R** software (through the **RStudio** interface)
  - I will provide **direct examples** of what you will need to do to complete HW (that also include sample results sections for your future reference!)