INTEGRATING ADVANCED DATA ANALYSIS TECHNIQUES AND COMPUTATIONAL SKILLS INTO EARLY COLLEGE CURRICULA

JAMIS J. PERRETT, PH.D.

DEPARTMENT OF STATISTICS

BRIGHAM YOUNG UNIVERSITY

Outline

- Focus on Student Preparation
- Intro Stats Example
- Data Science Ecosystem

FOCUS ON STUDENT PREPARATION

PREPARING STUDENT OF TODAY FOR THE CHALLENGES OF TOMORROW

Thought Questions

- Even if it is relevant today, will it still be relevant when students graduate?
- What will students remember 5 years from now?
- In which ways will students use statistics in their future careers?
- What are employers looking for?

Employers have indicated a need for data science students to have the following preparation:

- Python – coding languages
- Cloud Tools
  - Amazon Web Services (AWS)
  - Snowflake
  - DataBricks
- SQL – querying data
- Dashboarding – visualizing data

WHERE DO STUDENTS GET THEIR TECHNOLOGY SKILLS?

Many undergrads have no formal programming skills

CS & IS departments offer dedicated computing courses

How does the computing get implemented in the statistics course?

SOME IDEAS:

- For service courses, talk to the other departments, regularly. Collaborate on the course development
- For Intro courses,
  - Decide to use coding or point & click, depending on the need.
- For higher-level courses, use appropriate tools and stay current. Technology changes daily (DBMS list at 379 as of 1/5/2025, db-engines.com/en/ranking)
- Teach students how to use generative AI to aid in their learning process
  - Generative AI:
    - Maybe the best research assistant you will ever have
    - Can do the work of 25 programmers
  - Understand prompt engineering

SOME IDEAS:

- Implement programming in statistical methods courses
  - Use R/Python for computations
  - Introduce other tools like RStudio, VS Code, Anaconda, GitHub, Tableau as a means of facilitating the course
- Use more advanced technology in intro courses
  - RShiny app or webapp to facilitate/automate calculations
- Offer "data science" tools courses that introduce more advanced tools
  - Cloud, databases, SQL, Tableau, etc.
- Teach students how to use generative AI to aid in their learning process
  - Generative AI: Maybe the best research assistant you will ever have
  - Understand prompt engineering

INTRO STATS

BYU STAT 121 for various majors, no prerequisites, sometimes only statistics class students will take

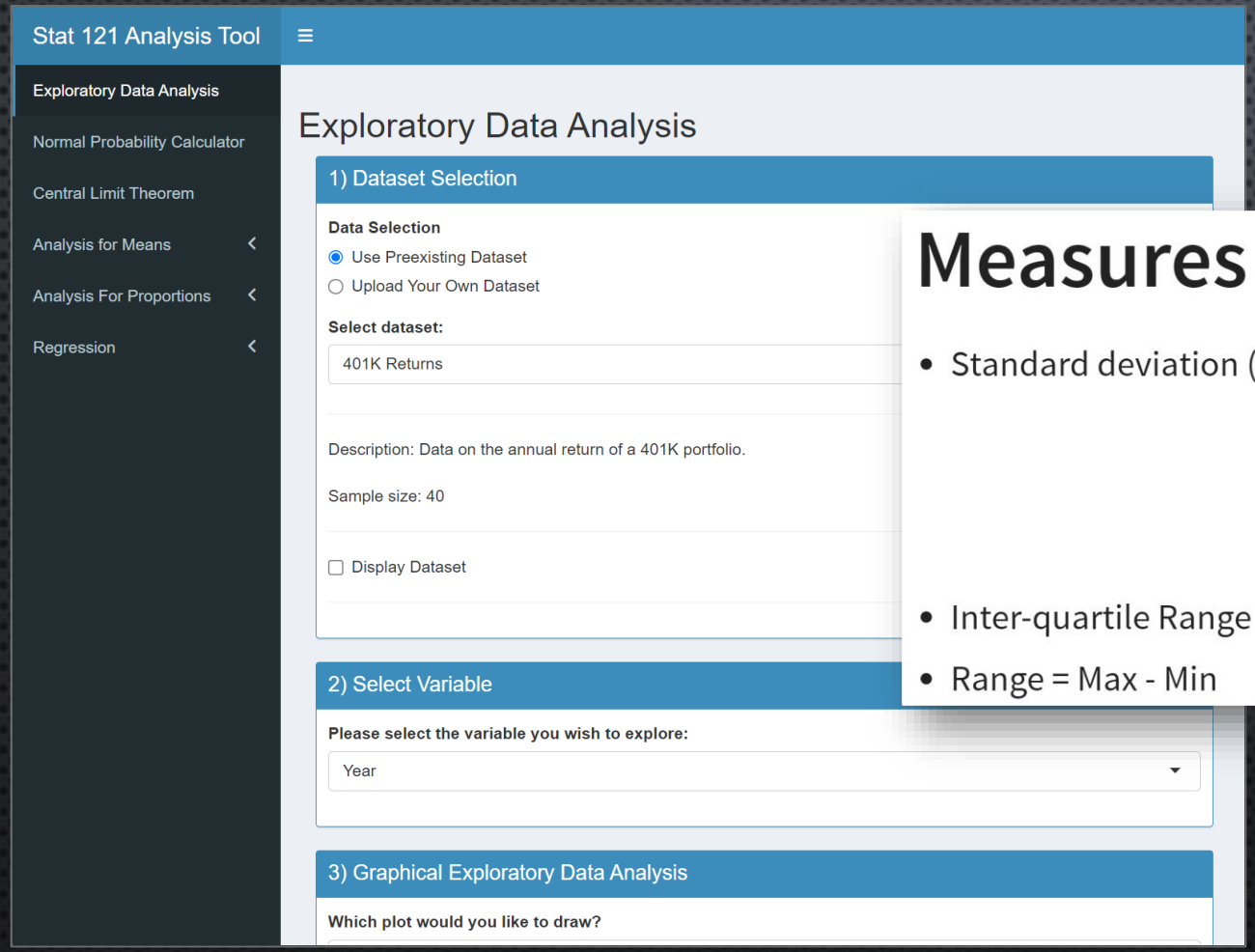
An Insightful Example

We start with real data

- Beginning of semester survey of 34 questions:
- College (categorical)
- Semesters (integer)
- In a relationship (Y/N)
- Mac or PC (Binary)
- Diet (categorical)
- Height and Mother's Height (continuous numeric)
- Streaming subscriptions (categorical)

When the data is about you, you are more interested in solving the problem

R-SHINY APP, POINT & CLICK TO AUTOMATE THE FORMULAS



Measures of Spread

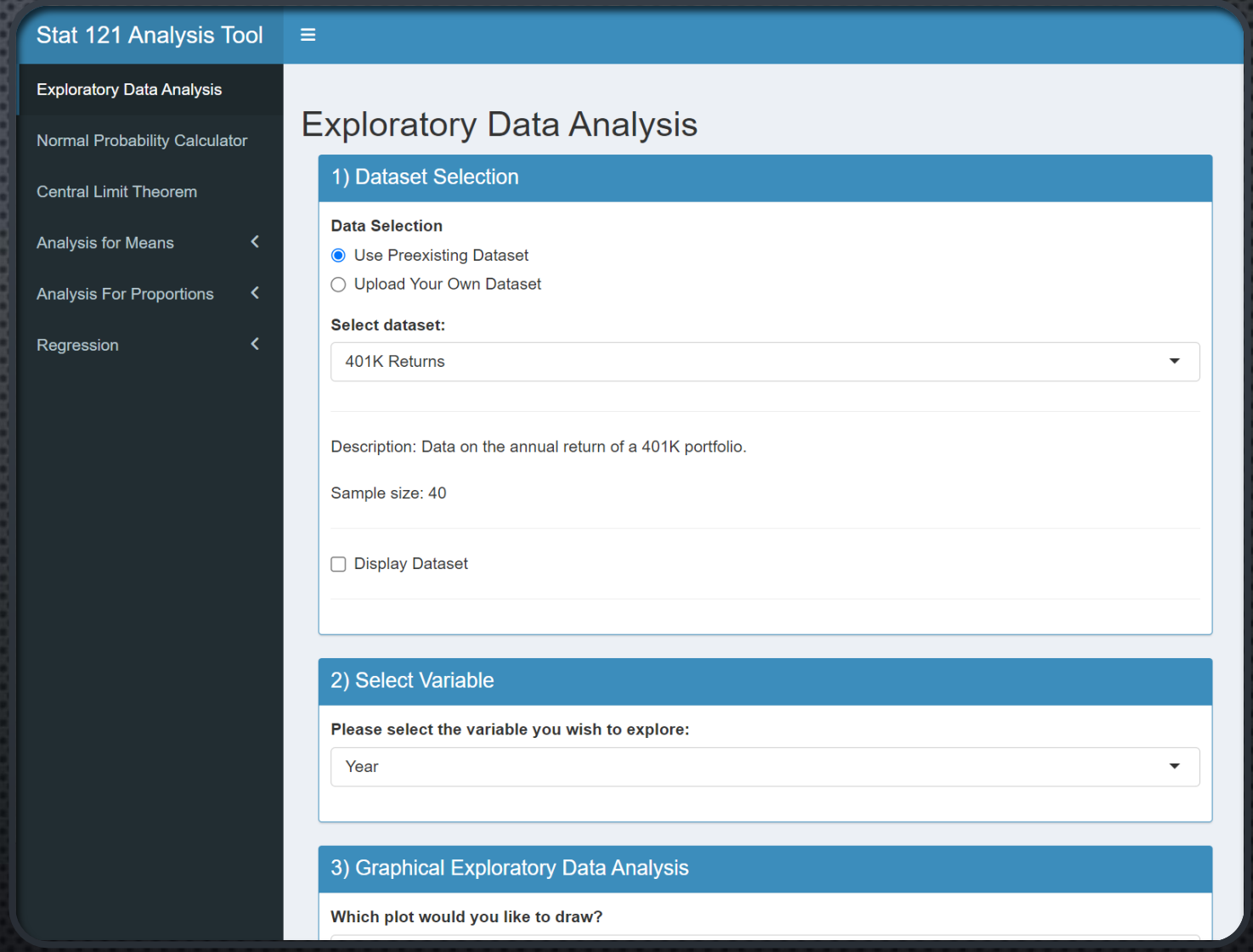
- Standard deviation (s): "average distance from the mean"

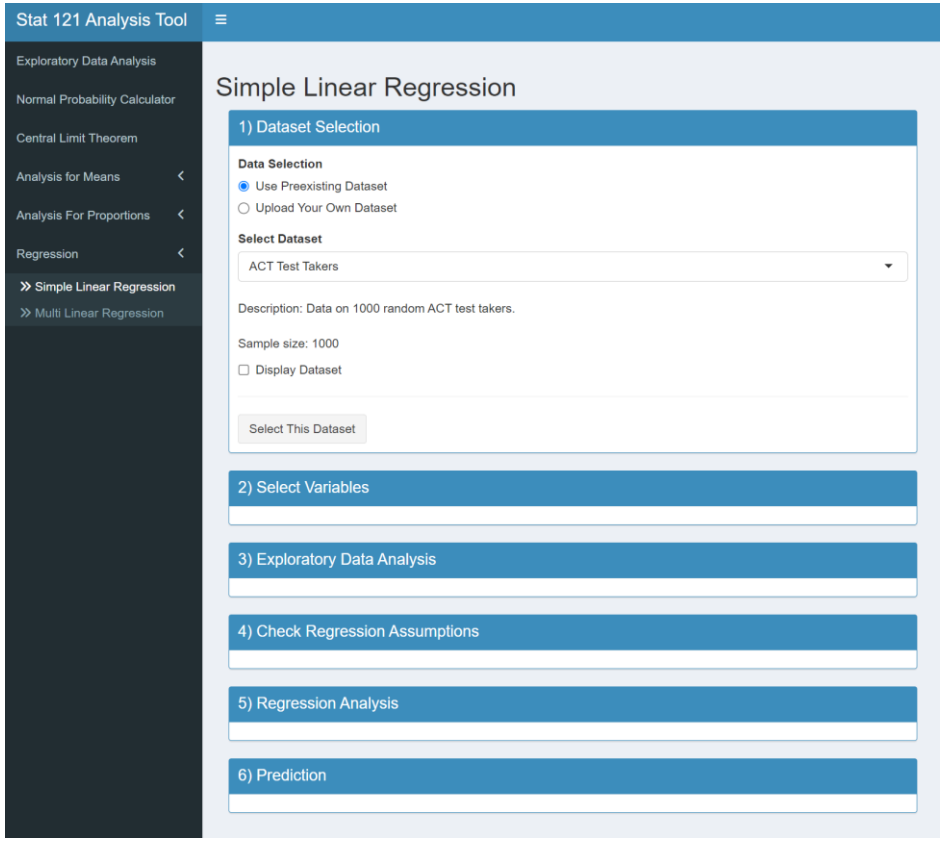
$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

- Inter-quartile Range = IQR = $Q_3 - Q_1$ = spread of the middle 50% of the data.
- Range = Max - Min

TOOLS CAN REINFORCE THE TOPICS COVERED IN THE COURSE (OUTLINE)

- Exploratory Data Analysis
- Normal Probability Calculator
- Central Limit Theorem
- Analysis for Means:
  - One Mean
  - Two Means
  - ANOVA
- Analysis for Proportions:
  - One Proportion
  - Two Proportions
  - Chi-Square
- Regression:
  - Simple Linear Regression
  - Multiple Linear Regression





TOOLS CAN REINFORCE PROPER STEPS FOR ANALYSIS

1. Data Selection
2. Select Variables
3. Exploratory Data Analysis
4. Check Regression Assumptions
5. Regression Analysis
6. Prediction

ASA COLLEGE GAISE RECOMMENDATIONS

1. Teach statistics and data science as iterative processes of gleaning insights from data to inform evidence-based decisions.
2. Emphasize effective written and oral communication of results from data, with attention to the scope and limitations of conclusions.
3. Focus on conceptual understanding rather than algebraic manipulation and formulas.
4. Integrate real data with a context and purpose throughout the course. Select data that are meaningful and engaging to the students.
5. Encourage multivariable thinking.
6. Incorporate software/apps to explore concepts and work with data.
7. Emphasize responsible and ethical conduct in the collection and use of data and in their analysis.
8. Employ evidence-based pedagogies that actively engage students in the learning process.
9. Use a variety of formative and summative assessments to improve teaching and learning.
10. Implement a course design that uses inclusive strategies to foster a sense of belonging.

DATA SCIENCE ECOSYSTEMS

Students planning to be data scientists,
Prerequisites: coding and intro stats,
Does not feed another class

An Insightful Example

BASIC OVERVIEW OF DIFFERENT DATA SCIENCE TOOLS

- Linux (SSH): Access remote data and work with permissions
- Containers: Using virtual environments
- **Databases and SQL: Understand how data are stored and how to access them**
- **AWS: Understand cloud environments**
- Dashboards: Interactive data visualizations on demand
- R & Python: Know how to code
- GitHub: Learn version control

PREPARING STUDENT OF TODAY FOR THE CHALLENGES OF TOMORROW

Repeat Slide

Thought Questions

- Even if it is relevant today, will it still be relevant when students graduate?
- What will students remember 5 years from now?
- In which ways will students use statistics in their future careers?
- What are employers looking for?

THANK YOU