Introduction to Data Analysis and Education Research using Stata

Norah Jones

2025-11-15

Table of contents

Preface

This book aims to teach you applied statistics in the context of education research. The goal of this book is to serve as an interactive introduction to quantitative analysis of educational data using Stata. The topics covered include descriptive statistics, correlation, statistical inference and hypothesis testing, and multiple regression. The content includes both a description of the concepts, and coding to implement the analyses using Stata.

There are already many books and other resources to learn statistics, or Stata. Why should you read this one (and why should we write it)? The aim of this book is to introduce statistics, from foundational concepts to sophisticated models, in an applied context, and focused on educational data.

This book is written using Quarto Book … The work in licensed under ….

Acknowledgements

My approach for much of the material presented comes from my training as a student in graduate school where I learned from Judith Singer, and serving as a teaching fellow, especially with Andrew Ho. I also built on the work of my colleagues in the education department John Papay and Matthew Kraft. Finally, I am thankful to the teaching assistants and students of EDUC 1230 and EDUC 2320 who have helped me refine teaching materials over the years. This book was inspired in part by Dr. Alice Paul's book Mastering Health Data Science Using R. The content for book.

1 Introduction to Stata

This chapter introduces you to Stata, how to install it and the different ways to use it for statistical analysis. Stata is a powerful tool, but it remains a tool. The statistical concepts you learn are more important, and you will likely have to change the tools you use throughout your career. As a Brown University affiliate, you can download and install the software following instructions on this page. There are many versions of Stata, and you should install the latest version available. The code included in this book will work for version 17 and those more recent. Stata / IC should be sufficient for most of the

1.1 Stata window components

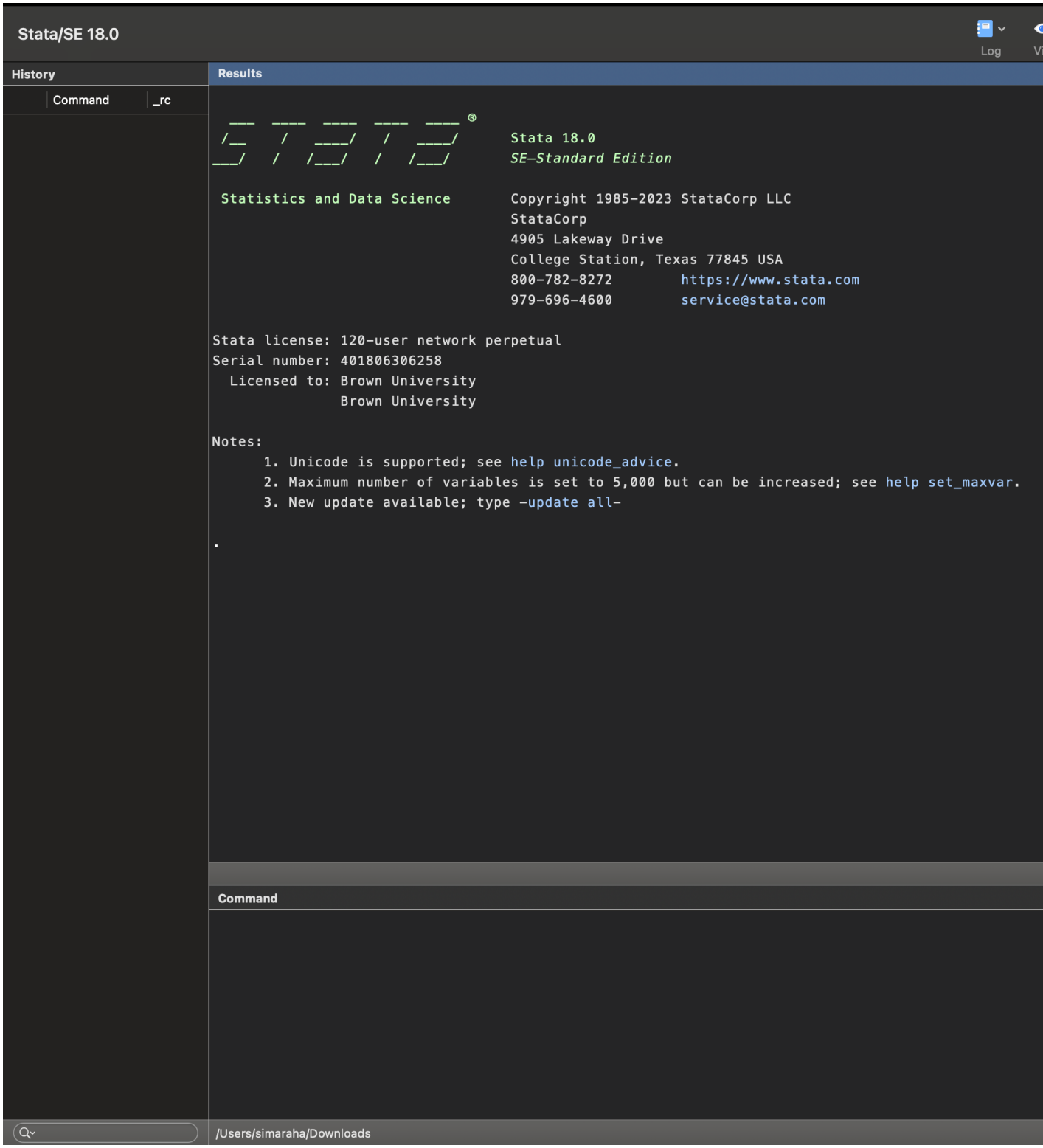

Figure 1.1: Stata Window

1.2 Stata syntax and Naming conventions

1.3 Using the command prompt

```
* ------------------------------------------------------------
* CLASS EXERCISE - UNIT 1
* Dataset: mcas_final_updated.dta
* Main task: practice basic Stata commands, descriptive
* statistics, and graphing distributions.
* ------------------------------------------------------------

* Load the dataset (adjust folder path if needed)
use "mcas_final_updated.dta", clear

* ------------------------------------------------------------
* 1. BASIC COMMANDS FOR VARIABLE: mathscore
* ------------------------------------------------------------

* Browse dataset (opens Data Browser)
browse

* Summary statistics: mean, sd, min, max
summarize mathscore

* Detailed summary: percentiles, variance, skewness, kurtosis, etc.
summarize mathscore, detail

* Stem-and-leaf plot
stem mathscore

* Boxplot
graph box mathscore, name(mathscorebox, replace)

* Histogram with frequencies
histogram mathscore, frequency name(mathscore, replace)

* ------------------------------------------------------------
* Saving graphs
* ------------------------------------------------------------

* Save graph in Stata's .gph format
graph save mathscore "dotmathscore.gph", replace
```

```
* Export histogram as .png
graph export "mathscore_hist.png", as(tif) width(2000) replace


* End of Part 1
* ------------------------------------------------------------

/* 2) Use the output from these commands to describe the distribution of mathscore in more de
How could you summarize what you see? Below, practice summarizing the univariate descriptive
statistics in no more than three sentences. (Remember to describe the following: unit of anal
central tendency, spread of the distribution, scale, skewness/symmetry, and atypical data poi
e.g., outliers).*/

/* 3) Now adapt this code for the ppe variable to produce univariate descriptive statistics a
histogram. Below, practice summarizing the univariate descriptive statistics in no more than
sentences.*/

* UNIVARIATE DESCRIPTIVE STATISTICS FOR PPE
* ------------------------------------------------------------

summarize ppe
summarize ppe, detail
stem ppe
graph box ppe, name(ppebox, replace)
histogram ppe, frequency name(ppehist, replace)

* Save graphs
graph save ppebox "doppebox.gph", replace
graph export "ppe_hist.png", as(tif) width(2000) replace
```

1.3.1 Make students use it as a calculator

1.4 Loading data

1.4.1 Introduce an educational dataset – maybe OECD, NCES, open dataset Working Directory

1.4.2 Stata code count, describe, labels, etc.

1.5 Installing Packages

1.5.1 Give some examples to do

1.6 Tips and resources

1.6.1 Give some links to pages

```
1 + 1 + 5
```

```
[1] 7
```

2 Introduction to Data and Univariate Descriptive Statistics

2.1 2.1 Types of Variables

Numerical Vs Categorical

Numerical: Continuous VS Discrete

Continuous: Interval VS Scale

Categorical: ordinal vs nonminal

Include definitions, examples

Load the data from Chapter 1

2.2 2.2 Mean, Median, Mode

2.3 2.3 Spread: Range, IQR, SD

2.4 2.4 Visualizing Univariate Data

Bar Chart

Histograms & Density Plots

3 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References