Data Visualization With Stata

Andy Grogan-Kaylor

4 Jun 2020

Introduction

- Stata is a powerful and intuitive data analysis program.
- Learning how to graph in Stata is an important part of learning how to use Stata. Yet, the default graphs in Stata can sometimes be less than optimal.
- This document is an introduction to (a) basic graphing ideas in Stata; and (b) some simple ways to make your Stata graphs look more professional.

What are Variables?

- By variables, I simply mean the columns of data that you have.
- For our purposes, you may think of variables as synonymous with questionnaire items, or columns of data.

Variable Types

- *categorical variables* represent unordered categories like *neighborhood*, or *religious affiliation*, or *place of residence*.
- *continuous variables* represent a continuous scale like a *mental health scale*, or a *measure of life expectancy*.

A Data Visualization Strategy

Once we have discerned the type of variable that have, there are two followup questions we may ask before deciding upon a chart strategy:

- Is our graph about **one thing at a time**?
  - How much of x is there?
  - What is the distribution of x?
- Is our graph about **two things at a time**?
  - What is the relationship of x and y?
  - How are x and y associated?

Figure 1: Norway Spruce and Larch Forest in Austrian Alps

Data Source

Image Source: https://ec.europa.eu/jrc/en/research-topic/forestry/qr-tree-project/norway-spruce

The data used in this example are derived from the R package *Functions and Datasets for "Forest Analytics with R"*.

According to the documentation, the source of these data are: "von Guttenberg's Norway spruce (Picea abies [L.] Karst) tree measurement data."

Figure 2: Old Tjikko, a 9,550 Year Old Norway Spruce in Sweden

The documentation goes on to further note that:

> "The data are measures from 107 trees. The trees were selected as being of average size from healthy and well stocked stands in the Alps."

```
use gutten.dta, clear
```

Variables

`site` Growth *quality* class of the tree's habitat. 5 levels.

`location` Distinguishes tree *location*. 7 levels.

`tree` An identifier for the tree within location.

`age.base` The tree age taken at ground level.

It might be best to use a centered age variable, centered at the grand mean of tree age:

```
egen ageMEAN = mean(age_base)

generate ageCENTERED = age_base - ageMEAN
```

`height` Tree height, m.

`dbh.cm` Tree diameter, cm.

`volume` Tree volume.

`age.bh` Tree age taken at 1.3 m.

`tree.ID` A factor uniquely identifying the tree.