Data Visualization using Stata

Iowa Social Research Center (ISRC) Workshop

Scott LaCombe

Department of Political Science
The University of Iowa
Iowa City, IA

October 4, 2018

Why Visualize Data?

- One can communicate information clearly and effectively via graphics
- Effective data visuals helps users analyze and reason with data
- Make complex data accessible, understandable and usable
- Display patterns and/or relationships in one's dataset
- One can visualize patterns and/or relationships with respect to discrete and/or continuous variables

`graph` *`type`* – Available Types

- Bar Graphs
- Box Plots
- Distribution Graphs
    - Histograms
    - Kernel Density Estimation Plots
- Dot Charts (Not Covered)
- Pie Charts (Not Covered)

Introduction

- Constructs bars used to visualize the distribution of a categorical variable
- Similar to a histogram
- Default is to construct a bar for each variable level

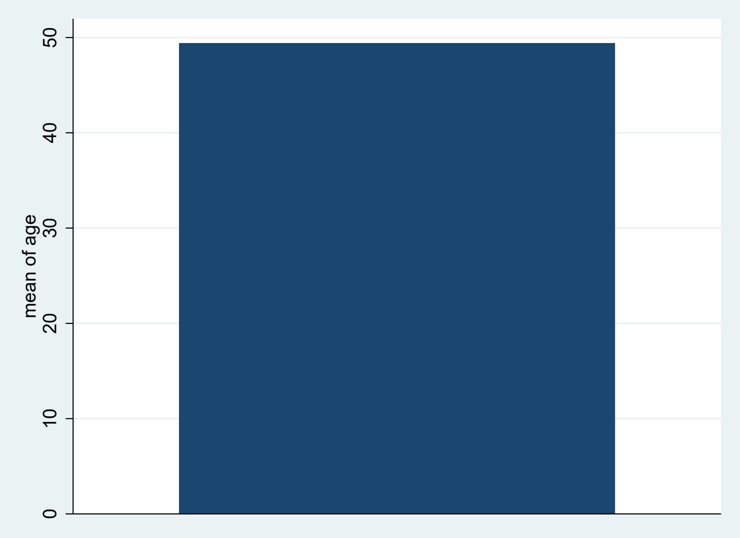
Syntax

- Basic: `graph bar (`*stat*`)` *yvars*, where *yvars* is a variable list
- Displays specified summary statistic for variable(s); default is the mean
- Other statistics include the median, count, various percentiles, etc.
- Can specify multiple (*stat*) *yvars*
- Can display summary statistic of specified variable based on levels of a categorical variable via the `over(`*varname*`)` option
- Advanced: `graph bar (`*stat*`)` *yvars*`, over(`*varname*`)`

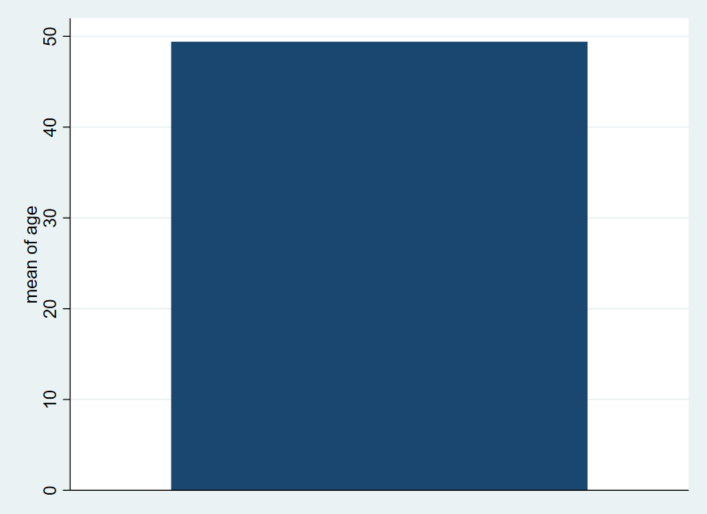
Syntax

- *yvars* is optional when `over(`*varname*`)` is specified (Stata 14 only)
- Acceptable syntax: `graph bar, over(`*varname*`)`
- Percentage is now treated as default statistic, calculated based on levels of *varname*
- Can change the statistic to `count`, which reports the frequency totals for each level of *varname*
- Replace `bar` with `hbar` to produce horizontal bar graph.
- See `help graph bar` for additional information

Bar Graph Example – PDF



Bar Graph Example – PNG



Introduction

- Displays a box and “whiskers” that visualizes the distribution of a continuous variable
- Box
  - Bordered at the 25th and 75th percentiles (Q1 and Q3)
  - An additional *median* line at the 50th percentile
- “Whiskers”
  - Lower Adjacent Value (LAV) – Smallest observation greater than or equal to the lower inner fence (LIF), which is $\text{Q1} - 1.5 \times \text{IQR}$, where $\text{IQR} = \text{Q3} - \text{Q1}$
  - Upper Adjacent Value (UAV) – Largest observation smaller than or equal to the upper inner fence (UIF), which is $\text{Q3} + 1.5 \times \text{IQR}$
- Any observation falling smaller (larger) than the adjacent values appears as dots

Syntax

- Basic: `graph box` *yvars* , where *yvars* is a variable list
- Can display box plots of specified variable(s) based on levels of a categorical variable via the `over(`*varname*`)` option
- Advanced: `graph box` *yvars*`, over(`*varname*`)`
- Replace `box` with `hbox` to produce horizontal box plot(s).
- See `help graph box` for additional information

Histograms

- A graph that shows the distribution of a variable that takes on many values (Acock 2014).
- Syntax: `histogram` *`varname`*`,` *`options`*
- Can be used for both discrete and continuous variables
- Use the command `help histogram` for more information

Kernel Density Estimation Plots

- Non-parametric method for estimating the PDF (PMF) of a random variable.
- Syntax: `kdensity` *`varname, options`*
- Can be used for both discrete and continuous variables
- Use the command `help kdensity` for more information

Introduction

- Used to display relationships between two numeric-type variables
- Represents over 30 different types of graphs, which can be grouped into multiple categories
- Easy to overlay `twoway`-type plots
  - Enclose graph type and variables in parentheses ()
  - Separate graphs via double vertical bars ||

Available Types Not Covered

- Area Plots
- Bar Plots
- Range Plots
- Regression Fits and Confidence Intervals
- Functions
- Contour Plots

Available Types Covered

- Scatterplots
- Line Plots
- Distribution Plots
    - Histogram
    - Kernel Density Plot

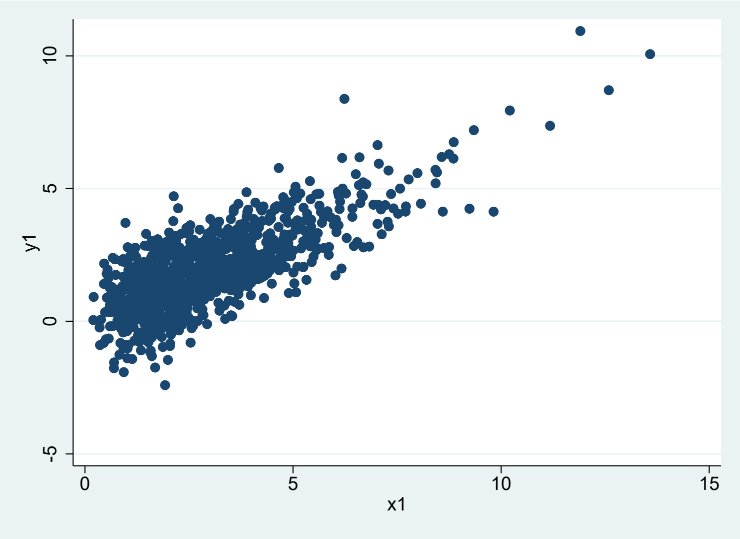
Introduction

- Utilize horizontal and vertical axes to plot data
- Communicates how much one variable is affected by another
- Visual representation of the correlation between two variables

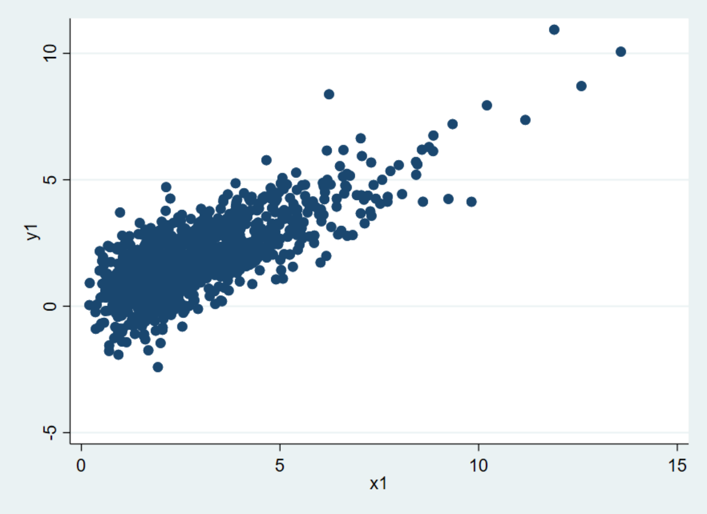
Syntax

- Basic: `scatter` *varlist*, where *varlist* is a variable list
- At least two variables need to be specified; last variable specified is treated as "independent" variable (located on the x-axis).
- Can generate scatterplots based on levels of a categorical variable via the `by(`*varname*`)` option
- Advanced: `scatter` *varlist*`, by(`*varname*`)`
- See `help scatter` for additional information

Scatterplot Example – PDF



Scatterplot Example – PNG



Introduction

- Shows frequency of data along a number line
- Similar to a scatterplot, except the points are connected
- Visual representation of a variable's trend

Syntax

- Basic: `twoway line` *`varlist`*, where *varlist* is a variable list
- At least two variables need to be specified; last variable specified is treated as “independent” variable (located on the x-axis).
- The default is to construct the graph based on the ordering of the dataset
- Either sort the dataset using the `sort` command, or use the `sort` option along with the `twoway line` command
- See `help twoway line` for additional information

twoway Version

- Same as hist and kdensity, except
  - Allows overlaying of a normal density or a kernel estimate of the density
  - If a density estimate is overlaid, it scales the density to reflect the scaling of the bars
- Basic Syntax
  - Histogram: `twoway histogram` *varname*
  - Kernel Density: `twoway kdensity` *varname*
- See `help histogram` and `help kdensity` for additional information

Graph Editor vs. Commands

- Two ways to edit Stata graphs: commands and the Graph editor
- Whenever possible, best to use commands to make changes to your graphs
- Situations can arise, however, where the graph editor is better suited than using the commands (e.g. adding objects, modifying objects)
- Changes made via the graph editor can be recorded, and applied to future graphs
- See *A Visual Guide to Stata Graphics, Third Edition*, pages 82-88, for a more detailed discussion

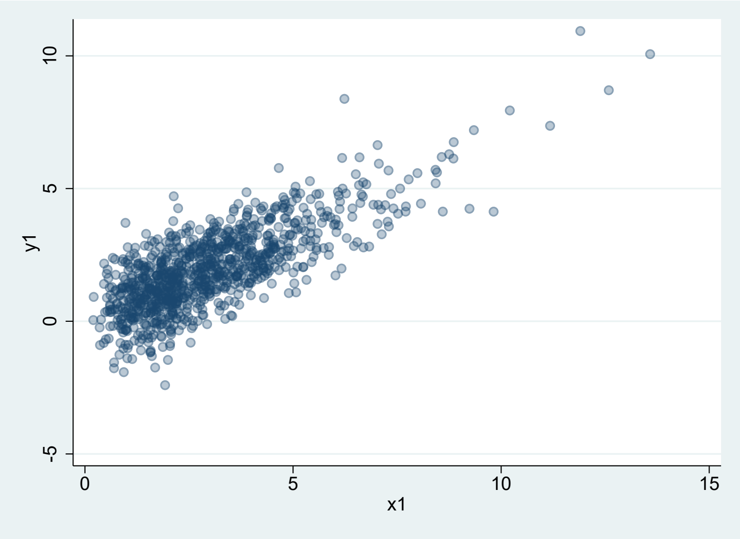
Scalable Vector Graphics (SVG)

- SVG images are scalable without image quality loss
- Used on webpages and EPUB ebook documents
- Compatible with modern desktop and mobile web browsers
- Editable with vector graphics applications and text editors (e.g. Adobe Illustrator)

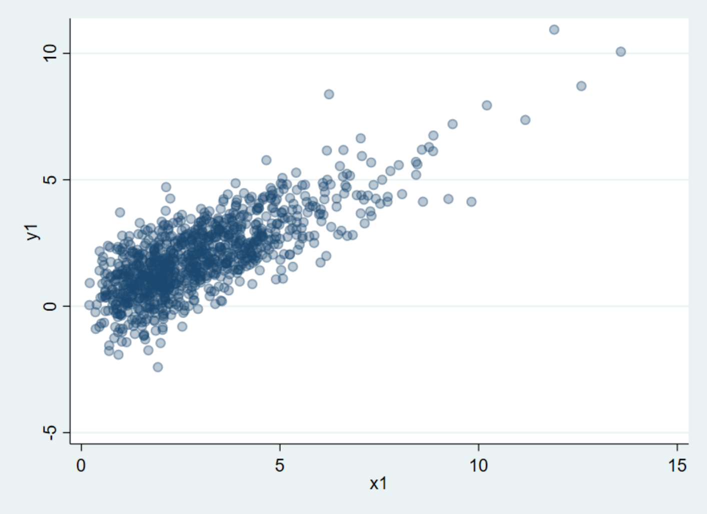
Graph Transparency

- Adjust color transparency in almost every element of a Stata graph
- Change percentage of opacity (Default: 100% opaque)
- See aspects of your data that weren't visible before
- Print graphs with transparency or export them to PDF, SVG, PNG, TIFF, or EMF
- NOTE: Only applicable in `twoway` graphs

Transparency Example – PDF



Transparency Example – PNG



Email: scott-lacombe@uiowa.edu
Any Questions?